



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/245030/>

Version: Accepted Version

Proceedings Paper:

Cheng, J.-K., Zhang, Z., Chen, R. et al. (2025) GeoUni: a unified model for generating geometry diagrams, problems and problem solutions. In: Gurrin, C., Schoeffmann, K., Zhang, M., Rossetto, L., Rudinac, S., Dang-Nguyen, D.-T., Cheng, W.-H., Chen, P. and Benois-Pineau, J., (eds.) MM '25: Proceedings of the 33rd ACM International Conference on Multimedia. MM '25: The 33rd ACM International Conference on Multimedia, 27-31 Oct 2025, Dublin, Ireland. ACM, pp. 3057-3066. ISBN: 9798400720352.

<https://doi.org/10.1145/3746027.3754965>

© 2025 The Authors. Except as otherwise noted, this author-accepted version of a conference paper published in MM '25: Proceedings of the 33rd ACM International Conference on Multimedia is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



GeoUni: A Unified Model for Generating Geometry Diagrams, Problems and Problem Solutions

Jo-Ku Cheng*

chengruogu@stu.pku.edu.cn
School of Mathematical Sciences,
Peking University
Beijing 100871, China

Zeren Zhang*

eric_zhang@stu.pku.edu.cn
School of Mathematical Sciences,
Peking University
Beijing 100871, China

Ran Chen

chenran@stu.pku.edu.cn
School of Mathematical Sciences,
Peking University
Beijing 100871, China

Jingyang Deng

jingyang@stu.pku.edu.cn
School of Mathematical Sciences,
Peking University
Beijing 100871, China

Ziran Qin

qinziran@sjtu.edu.cn
School of Electronic, Information and
Electrical Engineering, Shanghai Jiao
Tong University
Shanghai 200240, China

Jinwen Ma[†]

jwma@math.pku.edu.cn
School of Mathematical Sciences,
Peking University
Beijing 100871, China

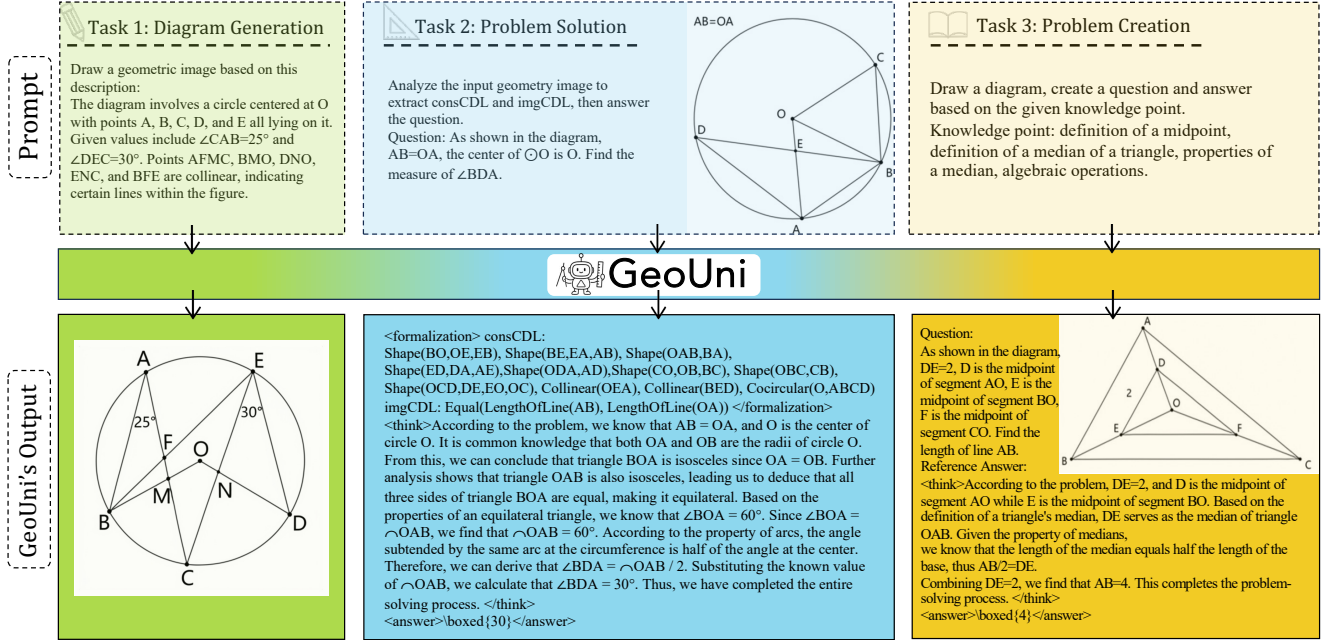


Figure 1: GeoUni can generate diagrams, solve problems and create new problems.

Abstract

We propose **GeoUni**¹, the first unified geometry expert model capable of generating problem solutions and diagrams within a single framework in a way that enables the creation of unique and individualized geometry problems. Traditionally, solving geometry problems and generating diagrams have been treated as separate tasks in machine learning, with no models successfully integrating both to support problem creation. However, we believe that

mastery in geometry requires frictionless integration of all of these skills, from solving problems to visualizing geometric relationships, and finally, crafting tailored problems. Our extensive experiments demonstrate that GeoUni, with only 1.5B parameters, achieves performance comparable to larger models such as DeepSeek-R1 with 671B parameters in geometric reasoning tasks. GeoUni also excels in generating precise geometric diagrams, surpassing both text-to-image models and unified models, including the GPT-4o image generation. Most importantly, GeoUni is the only model capable of successfully generating textual problems with matching diagrams

*Both authors contributed equally to this research.

[†]Corresponding author.

¹Our models are available at <https://github.com/chengruogu0915/GeoUni>.

based on specific knowledge points, thus offering a wider range of capabilities that extend beyond current models.

Keywords

Geometry Problem Solver, Multi-Modal Reasoning, Geometric Diagram Generation, Unified Model

1 Introduction

“If you want to master something, teach it.”
— Richard Feynman

Mastery of geometry involves not only the ability to solve problems, but also the skills to analyze and visualize geometric relationships, as well as the ability to teach and tutor others by creating new problems that challenge them individually. Automated geometry problem solving and diagram generation have traditionally been separate fields. The former emphasizes mathematical reasoning, while the latter focuses on accurately representing topological relationships and generating proper alphanumeric and angle annotations on diagrams. Existing models typically address either problem solving or diagram generation, and they fall short in creating new problems that cater to the specific learning goals of a student, which is a crucial aspect of individualized learning in mathematics. For this reason, these models are limited to simulating student performance and incapable of effectively assuming a tutor role.

This limitation stems from the lack of a unified framework for problem creation that combines multi-modal geometry understanding with the ability to generate diagrams, corresponding problem textual descriptions, and reference answers simultaneously. When a model can solve problems, generate diagrams, and create new questions based on specific knowledge points, it transitions from a passive tool to an active educator. This transition enables the model to offer an individualized learning experience, much like a tutor who tailors questions to challenge the learner, fostering a more interactive and engaging educational environment.

Although a unified model must ultimately overcome the challenge of integrating multiple tasks, but its priority remains executing each separate task with precision. First, solving geometry problems requires both abstract textual reasoning and precise visual understanding. Some existing geometry solver, such as AlphaGeometry [34] and FGPS [42], excel in geometric reasoning tasks but rely exclusively on textual inputs. There are also multi-modal models attempting to solve geometry problems [12, 43]. However, these models are limited to understanding diagrams and lack the ability to generate them.

Second, current diagram generation tools like GeoGebra [15] provide interactive graphical interfaces that rely heavily on manual user input through mouse interaction. Traditional text-to-image models, such as diffusion models [29] are primarily trained on natural images and thus struggle to generate accurate geometric diagrams. Even recent unified models like GPT-4o [25], despite significant advancements in general image generation, including textual content, still fall short in accurately plotting precise geometric diagrams.

To address these challenges, we propose **GeoUni**, the **first unified model** designed to integrate generating geometry problems,

diagrams, and problem solutions seamlessly. As illustrated in Fig. 1, GeoUni demonstrates strong performance across text-to-diagram generation, geometric reasoning, and geometry problem creation tasks. Our model performance in geometric diagram generation surpasses existing models across various metrics. Additionally, GeoUni achieves geometric reasoning performance comparable to much larger models, accomplishing this with only 1.5B parameters across three datasets in both multiple choice and open-ended question modes. Furthermore, GeoUni demonstrates a unique capability in geometry problem generation that goes beyond the limitations of existing models.

To facilitate better representation and tokenization of geometric diagrams, we propose **Geo-MAGVIT** designed to capture detailed geometric structures and reconstruct diagrams accurately. We observe that prior work, such as MagicGeo [35], evaluates diagram quality using the CLIP score. However, this metric is unsuitable for geometric diagrams, as CLIP is pre-trained on natural images and fails to capture the structural and symbolic characteristics. For a comprehensive evaluation of diagram quality, we introduce two new metrics: the **Geometry Semantic Matching Scores (GSMSs)**, which evaluates the alignment of geometry semantics, and the **Geometry Pixel Matching Score (GPMS)**, which assesses pixel-level fidelity. Additionally, we propose **Geo-Reasoning-Adapter**, which effectively leverages LoRA and GRPO to significantly enhance the model’s reasoning ability for geometry problem solving without affecting its diagram generation capability.

The main contributions of this work are:

- We propose the first unified multi-modal geometry expert model, **GeoUni**, capable of solving geometry problems, generating precise geometric diagrams using both formal and natural language, and creating geometry problems based on knowledge points. All three tasks are supported in both English and Chinese.
- We propose **Geo-MAGVIT**, a module specifically designed for the tokenization of geometric diagrams. By introducing topo-structural awareness loss and text region loss, it significantly improves the precision of geometry structure and text reconstruction.
- We innovatively combine GRPO and LoRA to train the **Geo-Reasoning-Adapter**, which effectively boosts geometric reasoning capability and seamlessly integrates into the unified model architecture.
- We establish a novel diagram generation evaluation metrics, which includes the **Geometry Semantic Matching Scores (GSMSs)** and **Geometry Pixel Matching Score (GPMS)** to comprehensively evaluate the diagram generation task.

2 Related Work

2.1 Unified Model

Multi-modal Large Language Models (MLLMs) [1, 18, 24] are primarily designed to process images or videos as input and generate only text as output. This limitation has driven the development of unified models that integrate both multi-modal understanding and generation [7, 14, 32]. One of the first unified models, UNIFIED-IO [19], integrates text and image encoding into discrete tokens,

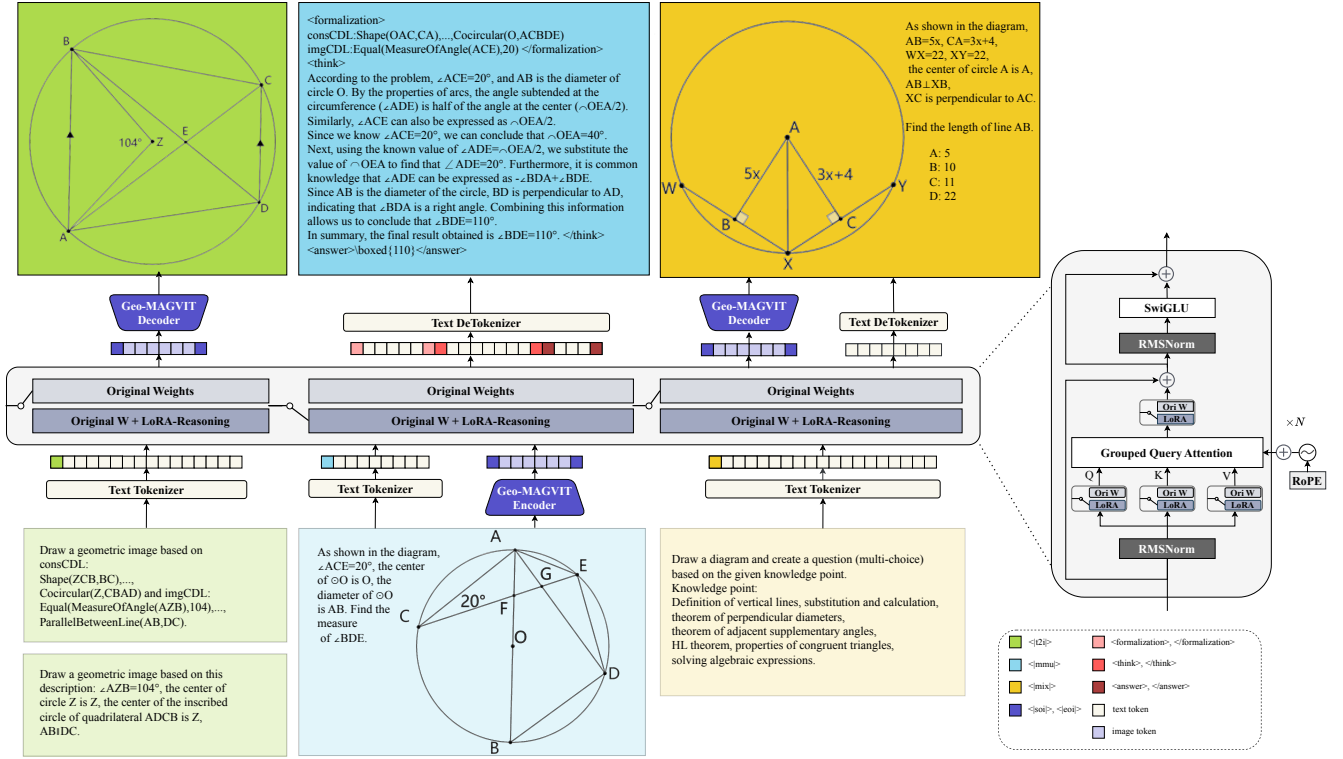


Figure 2: Overview of GeoUni

enabling unified processing across multiple modalities. The Emu series [31, 36] further unifies video, image, and text modeling within a next token prediction framework, while SEED-LLaMA [13] and Show-o [38] introduce techniques such as novel image tokenization and discrete diffusion modeling for improved performance. However, these models often struggle with geometric diagram generation, as diagrams present unique structural challenges not well addressed by standard image generation techniques.

2.2 MLLM-based Geometry Problem Solver

MLLM-based geometry problem solvers fall into two categories: those generating formal language programs requiring symbolic execution [20, 40], and those producing directly readable natural language answers.

The first category such as GeoX [37] introduces unimodal pre-training, geometry-language alignment, and end-to-end instruction tuning to train an MLLM capable of generating formal language reasoning steps. The second category, exemplified by G-LLaVA [12], follows the LLaVA training strategy, leveraging GPT-3.5 to construct the multi-modal geometry dataset, Geo170K. This dataset focuses on geometric cross-modal alignment and geometric instruction tuning to generate human-readable solutions. To address the mismatch between text descriptions and diagrams in Geo170K, DFE-GPS [43] incorporates geometric formal language into diagram descriptions and creates a large-scale synthetic dataset SynthGeo228K to better

train the Diagram Formalizer, enhancing the model’s ability to understand and generate accurate geometric representations. Despite these efforts, these models can understand diagrams, but still can not generate geometric diagrams.

2.3 Automated Geometric Diagram Generation

The Geometry Model Builder [17] introduces the Geometry Model-Building Language (GMBL) to represent diagrams, treating the diagram creation process as a numerical optimization problem solved through gradient descent. Other approaches leverage natural language and LLMs to complete the process. GeoGPT4V [4] utilizes GPT-4 to generate Wolfram code, which is executed to produce the diagram. And MagicGeo [35] prompts an LLM to formalize the diagram’s description by encoding coordinate points and geometric constraints, which are then passed to a solver to find precise coordinate solutions. The LLM subsequently generates TikZ code to render the final diagram. However, all these models rely on generating formal language representations or code for rendering engines or solvers to construct diagrams, rather than adopting an end-to-end approach that directly generates diagrams from text.

3 Preliminaries

3.1 Low-Rank Adaptation (LoRA)

LoRA[16] is widely used for fine-tuning LLMs in various downstream tasks, as it preserves the performance of the base model

while mitigating the issue of forgetting[3]. The implementation of LoRA is straightforward: instead of updating the full weight matrix $W \in \mathbb{R}^{m \times n}$, it introduces two low-rank matrices, $A \in \mathbb{R}^{r \times n}$ and $B \in \mathbb{R}^{m \times r}$, where $r \ll \min(m, n)$. After training the low-rank matrices A and B , the target weight is determined by the following expression:

$$W_{target} = W_{base} + \Delta W = W_{base} + BA. \quad (1)$$

3.2 Group Relative Policy Optimization (GRPO)

GRPO [9, 30] reduces the training costs of reinforcement learning (RL) by eliminating the need for a value model in the training loop. It utilizes the sampled outputs $\{o_1, o_2, \dots, o_G\}$ from the policy model to compute the corresponding rewards $\{r_1, r_2, \dots, r_G\}$, which are then used to compute the group normalized score as a relative advantage estimate:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (2)$$

Then the policy model is optimized using the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E} \left[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q) \right] \\ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left[\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t} - \beta \mathbb{D}_{KL}(\pi_{\theta} \parallel \pi_{ref}) \right]. \quad (3)$$

4 Methodology

4.1 Overview

Our model, GeoUni, needs to address several challenges. First, because the unified model’s vision tokenizer is trained on general images, it faces the same issues identified by [37, 43]. This causes it to be ineffective in tokenizing geometric diagrams, and limits its ability to accurately reconstruct and generate them. Second, effectively integrating the three tasks into a unified training framework remains a non-trivial challenge. Finally, another key difficulty is to enhance the model’s reasoning capabilities without compromising its diagram generation ability. To address these issues, the training pipeline is organized into three stages, each with its own focus:

- **Diagram Tokenization Pretraining.** We propose **Geo-MAGVIT** to improve the tokenization of geometric diagrams. Building on MAGVIT[22], we introduce geometric topo-structural awareness loss and text region loss to better reconstruct the topological structure and the text within the diagrams.
- **Multi-Task Instruction Tuning.** To achieve the geometry expert unified model, we propose the **Diagram Formalization Unified Prompting** method in multi-task instruction tuning for text-to-diagram generation, problem solving, and problem generation, achieving next-token prediction training. This training phase equips GeoUni with the capability to accurately generate geometric diagrams, solve basic geometry problems, and generate problems based on knowledge points.
- **Reasoning Enhancement.** We combine LoRA and GRPO to train the **Geo-Reasoning-Adapter**, which significantly

improves the model’s geometric reasoning ability while preserving its precise geometric diagram generation capability.

4.2 Diagram Tokenization Pretraining

Following MAGVIT [22], we pre-train the Geo-MAGVIT on a geometry dataset consisting of approximately 200K diagrams.

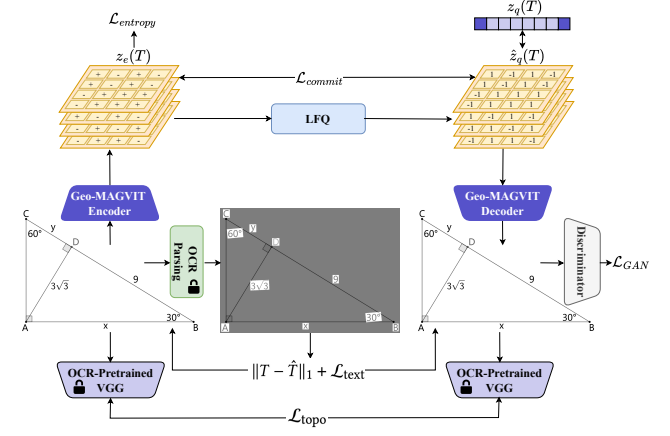


Figure 3: Overview of Geo-MAGVIT

Given a diagram $T \in \mathbb{R}^{H \times W \times 3}$, we extract the representation after the Geo-MAGVIT Encoder, denoted as $z_e(T) \in \mathbb{R}^{H' \times W' \times \log(C)}$. We flatten it along the spatial dimension as $z_e(T) = \{z_e^i(T)\}_{i=1}^{H'W'}$. Given a feature vector $z_e^i(T) \in \mathbb{R}^{\log(C)}$, we apply Look-Up-Free Quantization (LFQ), where the codebook becomes an integer set $\mathbb{C} = \prod_{j=1}^{\log(C)} \{-1, 1\}$, and the latent space of each vector is decomposed as the Cartesian product. LFQ quantizes it according to the following equation:

$$\hat{z}_q^i(T) = \text{sign}(z_e^i(T)) = -[z_e^i(T) \leq 0] + [z_e^i(T) > 0]. \quad (4)$$

This gives us $\hat{z}_q(T) = \{\hat{z}_q^i(T)\}_{i=1}^{H'W'}$, and the reconstructed diagram is then obtained as:

$$\hat{T} = \mathcal{D}(z_e(T) + \text{sg}[\hat{z}_q(T) - z_e(T)]). \quad (5)$$

Additionally, we can obtain the specific image token index representation $z_q(T)$ from $\hat{z}_q(T)$ as follows:

$$z_q^i(T) = \sum_{j=1}^{\log(C)} 2^{j-2} [\hat{z}_q^{i,j}(T) + 1], \quad i = 1, \dots, H'W'. \quad (6)$$

The training incorporates multiple loss functions, including GAN loss, reconstruction loss, commit loss, and entropy loss, which are optimized as a weighted sum below:

$$\mathcal{L}_{\text{Geo-MAGVIT}} = \mathcal{L}_{\text{GAN}} + \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} + \lambda_{\text{commit}} \cdot \mathcal{L}_{\text{commit}} + \lambda_{\text{entropy}} \cdot \mathcal{L}_{\text{entropy}}. \quad (7)$$

We observe that MAGVIT encounters difficulties when reconstructing letters, numeric symbols and topological structures in the

diagrams. To address this issue, we redesign the reconstruction loss $\mathcal{L}_{\text{rec}}$ by incorporating both $\mathcal{L}_{\text{topo}}$ and $\mathcal{L}_{\text{text}}$:

$$\mathcal{L}_{\text{rec}} = \|T - \hat{T}\|_1 + \mathcal{L}_{\text{topo}} + \mathcal{L}_{\text{text}}. \quad (8)$$

The topo-perceptual loss is to enhance the precision of the geometric topological structure in the generated diagrams. For implementation, we use the loss between features pre-trained on the VGG model for document OCR tasks [28], and it is formulated as follows:

$$\mathcal{L}_{\text{topo}} = \sum_{i=1}^M \|F_{\text{vgg}}^{(i)}(T) - F_{\text{vgg}}^{(i)}(\hat{T})\|_1. \quad (9)$$

We also introduce $\mathcal{L}_{\text{text}}$ to improve the accuracy of textual reconstruction. We apply the OCR tool [26] to generate bounding boxes for critical regions, such as endpoint labels and length/angle annotations on line segments within the diagrams. The design of $\mathcal{L}_{\text{text}}$ is as follows:

$$\mathcal{L}_{\text{text}} = \|M \odot (T - \hat{T})\|_1. \quad (10)$$

Besides, the original MAGVIT entropy loss is a convex function, and therefore always non-positive. We can prove that the minimum value of the entropy loss is $-\log(C)$ (proof in Appendix), hence we add a $\log(C)$ term to guarantee training stability:

$$\mathcal{L}_{\text{entropy}} = \mathbb{E}[H[f(z_e(T))]] - H[\mathbb{E}[f(z_e(T))]] + \log(C). \quad (11)$$

4.3 Multi-Task Instruction Tuning

We initialize GeoUni using the weights of a pre-trained LLM and treat the multi-task instruction tuning as next-token prediction.

4.3.1 Diagram Formalization Unified Prompting. To perform multi-task instruction tuning, we design the Diagram Formalization Unified Prompting to organize various types of data into a structured format. We pre-define three special tokens: $\langle |t2i| \rangle$, $\langle |mmu| \rangle$, and $\langle |mixing| \rangle$, which represent the three tasks: text-to-diagram, problem-solving, and problem-generation. Additionally, $\langle |soi| \rangle$ and $\langle |eoi| \rangle$ are special tokens used to mark the start and end of discrete diagram tokens. $\langle |formalization| \rangle$ and $\langle |/formalization| \rangle$ are used to mark the beginning and end of the formalized description of the diagram. $\langle |think| \rangle$ and $\langle |/think| \rangle$ denote the start and end of the reasoning process in solving geometry problems, while $\langle |answer| \rangle$ and $\langle |/answer| \rangle$ mark the final answer. As shown in Figure 4, by adding different task tokens as the start of the sequence to distinguish different tasks, all data is converted into a 1D sequence of tokens.

4.3.2 Training Objectives. After processing with Geo-MAGVIT, we obtain the image tokens $\mathbf{d} = \{d_1, d_2, \dots, d_N\}$. The instruction and response text are also tokenized as $\mathbf{t} = \{t_1, t_2, \dots, t_M\}$ and $\mathbf{r} = \{r_1, r_2, \dots, r_K\}$, respectively. To perform unified next token prediction training, we employ three training objectives for the three different tasks.

For the text-to-diagram task, we minimize the negative log-likelihood of the diagram tokens based on the instructions:

$$\mathcal{L}_{T2D} = \mathbb{E}_{(\mathbf{d}, \mathbf{t}) \sim D_{T2D}} \left[- \sum_{i=1}^N \log p_{\theta}(d_i | \mathbf{t}, d_1, \dots, d_{i-1}) \right]. \quad (12)$$

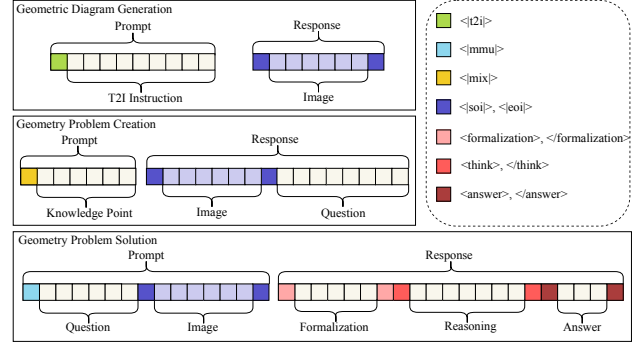


Figure 4: Diagram Formalization Unified Prompting

For the geometry problem solution task, this task is a standard multi-modal understanding task, where the response answer is generated based on the provided problem text and diagram:

$$\mathcal{L}_{MMU} = \mathbb{E}_{(\mathbf{d}, \mathbf{t}, \mathbf{r}) \sim D_{MMU}} \left[- \sum_{i=1}^K \log p_{\theta}(r_i | \mathbf{t}, \mathbf{d}, r_1, \dots, r_{i-1}) \right]. \quad (13)$$

For the geometry problem generation task, the process first generates the diagram and then the text, which involves mixing both text and image tokens. The loss function is defined as:

$$\begin{aligned} \mathcal{L}_{MIX} = & \mathbb{E}_{(\mathbf{d}, \mathbf{t}, \mathbf{r}) \sim D_{MIX}} \left[- \sum_{i=1}^N \log p_{\theta}(d_i | \mathbf{t}, d_1, d_2, \dots, d_{i-1}) \right. \\ & \left. - \sum_{j=1}^K \log p_{\theta}(r_j | \mathbf{t}, \mathbf{d}, r_1, \dots, r_{j-1}) \right]. \end{aligned} \quad (14)$$

For joint training of the three tasks within a single framework, the total loss is defined as a weighted loss:

$$\mathcal{L}_{GeoUni} = \lambda_{T2D} \cdot \mathcal{L}_{T2D} + \lambda_{MMU} \cdot \mathcal{L}_{MMU} + \lambda_{MIX} \cdot \mathcal{L}_{MIX}. \quad (15)$$

4.4 Reasoning Enhancement

Modality-specific functionality requires fine-tuning the base language model, which can hinder its original capabilities [24]. Different tasks, like different modalities, present unique challenges. To address this, we employ GRPO and LoRA to fine-tune the reasoning adapter, enhancing reasoning performance without compromising the diagram generation capability of the instruction fine-tuned model. We redesign the reward function to better facilitate geometric reasoning tasks. The total reward function is defined as the sum of three components: format reward, formalization reward, and accuracy reward.

Format Reward We encourage the model to structure its responses using a pre-defined format: $\langle |formalization| \rangle$ for diagram formalization, $\langle |think| \rangle$ for the reasoning process, and $\langle |answer| \rangle$ for the final answer. This reward is given a score of 1.0 if the response follows the above structure.

Formalization Reward Formalizing the diagram before reasoning helps the model better understand its geometric relationship. We supervise this process using a formalization score based on the Levenshtein distance between the predicted and ground truth consCDL

and imgCDL, denoted as d_{consCDL} and d_{imgCDL} , respectively. We define the individual scores as follows:

$$S_{\text{consCDL}} = 1 - \frac{d_{\text{consCDL}}}{\max(|y_{\text{consCDL}}^*|, 1)}, \quad (16)$$

$$S_{\text{imgCDL}} = 1 - \frac{d_{\text{imgCDL}}}{\max(|y_{\text{imgCDL}}^*|, 1)}. \quad (17)$$

The formalization reward is computed as the average of these two scores:

$$R_{\text{formal}}(y, y^*) = \max\left(0, \frac{S_{\text{consCDL}} + S_{\text{imgCDL}}}{2}\right). \quad (18)$$

Accuracy Reward After performing formalization, which consists of structuring consCDL and imgCDL, the model proceeds to generate the reasoning process in natural language and the final answer. For both four-option multiple choice and open-ended questions, accuracy is measured based on whether the model’s output exactly matches the standard answer.

5 Experiments

5.1 Datasets

We train GeoUni on Formalgeo7K [41] and SynthGeo228K [43]. Each Formalgeo7K sample includes a bilingual description, a diagram, consCDL capturing topological relations, imgCDL for other geometric constraints, and a formalSSS symbolic solution translated into natural language. Considering the unique characteristics of geometry, we further design task-specific data augmentation strategies tailored for each task. For more details, please refer to the Appendix.

5.2 Implementation Details

The Geo-MAGVIT Encoder downsamples the input image resolution from 512×512 to 256 tokens and is trained for 50 epochs with a batch size of 16. For the base LLM model, we adopt DeepSeek-R1-Distill-Qwen-1.5B [9]. Multi-task instruction tuning is conducted for 50K steps with a batch size of 16. The Geo-Reasoning-Adapter is trained using LoRA with a rank of 256, applied to the q, k, v, and o projection modules. Training is performed with a batch size of 4 and a gradient accumulation step of 4. GRPO samples 8 responses per question and is trained for 4 epochs. All models are trained on 4 NVIDIA A800 (80GB) GPUs.

5.3 Diagram Reconstruction

5.3.1 Metrics. To evaluate the quality of diagram reconstruction, we design two types of metrics. One evaluates semantic accuracy in formal language, while the other focuses on pixel-level accuracy.

Geometry Semantic Matching Scores (GSMSs) We use the geometric parser from [45] to translate the generated diagrams into these two CDL formats. Two precision metrics are proposed for evaluation: **Average Accuracy (AA)**, representing the average percentage of matched statements after transformation, and **Perfect Accuracy (PA)**, indicating the proportion of completely correct statements. We compute AA and PA separately for consCDL (C-AA, C-PA) and imgCDL (I-AA, I-PA), as well as a combined Perfect

Accuracy (CI-PA) that counts diagrams perfectly matching both CDLs.

Geometry Pixel Matching Score (GPMS) Geometric diagrams are characteristically monochromatic and highly structured; only the black pixel regions encode geometric meaning, while the white background carries no task-relevant information. We define the geometric pixel sets as F_{Gt} and F_{Rec} where the pixels are black in the reference and reconstructed diagrams, respectively. The GPMS is then computed as:

$$\text{GPMS} = 2 \times \frac{|F_{Gt} \cap F_{Rec}|}{|F_{Gt}| + |F_{Rec}|}. \quad (19)$$

5.3.2 Results. Table 1 provides a comparative analysis of diagram reconstruction performance across various image tokenizers, including MAGVIT [22] and our proposed Geo-MAGVIT. Geo-MAGVIT achieves superior results in both GSMSs and GPMS, demonstrating its strong ability to reconstruct diagrams with high fidelity, which is essential for the text-to-diagram task. While GPMS evaluates the accuracy at the pixel level, GSMSs focus on preserving the semantics of geometry. UniTok [23] and QLIP [44] perform reasonably well on GSMSs, indicating that they can preserve the basic shapes of geometric diagrams. However, they fail to achieve accurate one-to-one reconstructions, often missing fine-grained details.

Table 1: Geometric Diagram Reconstruction Performance Comparison of Various Models Across Different Matrices

Model	C-AA	C-PA	I-AA	I-PA	CI-PA	GPMS
UniTok[23]	81.43	54.29	69.95	48.94	31.44	39.65
QLIP[44]	77.82	50.63	68.73	48.10	30.06	52.87
MAGVIT[22]	80.78	52.10	66.4	43.52	27.52	86.73
Geo-MAGVIT(Ours)	83.10	55.05	75.98	55.71	35.24	91.32

5.4 Text-To-Diagram

5.4.1 Metrics. In the text-to-diagram generation task, evaluating visual outputs is challenging due to the absence of pixel-level ground truth, making traditional image similarity metrics inapplicable. To address this, we adopt a symbolic evaluation approach. Specifically, we parse the generated diagrams into consCDL and imgCDL formats using a geometry parser as before, and compute BLEU-4 scores against the reference CDLs representations. This metric reflects structural fidelity without relying on pixel-level alignment and is effective in detecting inconsistencies in predicate composition, object relationships, and the symbolic structure of the diagrams.

5.4.2 Results. In Table 2, we show a comparative analysis of text-to-diagram generation performance across unified models, text-to-image (T2I) models, and our proposed GeoUni, under three types of prompts: natural language captions, formalized CDL descriptions, and GPT-rewritten instructions, in both English and Chinese. GeoUni consistently achieves the highest scores across all settings, significantly outperforming baselines in both consCDL and imgCDL BLEU-4 metrics. It is important to note that BLEU-4 is a relatively

Table 2: Text-To-Diagram Performance Comparison of Various Models Across Different Matrices

Model	Construction/EN			Image/EN			Construction/CN			Image/CN		
	Caption	CDL	GPT	Caption	CDL	GPT	Caption	CDL	GPT	Caption	CDL	GPT
Show-o[38]	29.39	25.01	24.84	34.51	24.58	34.01	32.67	29.82	35.63	33.80	31.33	34.00
Janus-Pro-7B[7]	14.55	17.72	20.85	30.22	29.7	33.02	34.85	35.47	34.69	29.09	20.56	26.55
Anole-7B[8]	16.84	18.68	13.63	24.11	30.73	23.73	12.42	19.55	14.25	34.01	28.59	34.40
Emu3[36]	22.06	18.76	16.00	20.89	19.88	19.90	21.20	25.25	18.78	20.96	22.65	19.75
Unified-IO[19]	32.27	29.37	31.63	17.84	16.41	16.74	37.62	33.81	32.51	27.95	19.43	29.44
SEED-X[14]	34.20	31.78	27.84	14.80	10.47	12.30	30.78	31.40	27.40	15.60	11.37	14.89
PixArt-Σ[5]	18.51	17.94	17.14	19.35	19.44	15.39	28.55	23.67	28.29	17.89	21.40	16.45
SD-v1.5[29]	21.69	19.51	23.06	22.46	20.71	22.26	25.03	25.14	27.30	14.11	13.34	16.72
SDXL-Turbo[11]	12.94	11.12	14.76	13.86	15.93	11.76	30.36	17.86	29.62	19.52	10.84	18.38
DALL-E-2[27]	24.38	19.75	18.30	28.29	24.74	19.27	22.50	19.56	20.78	20.31	21.36	17.47
GeoUni-1.5B(Ours)	73.00	73.43	72.41	78.46	79.65	77.53	73.72	73.00	72.77	79.40	79.54	77.82

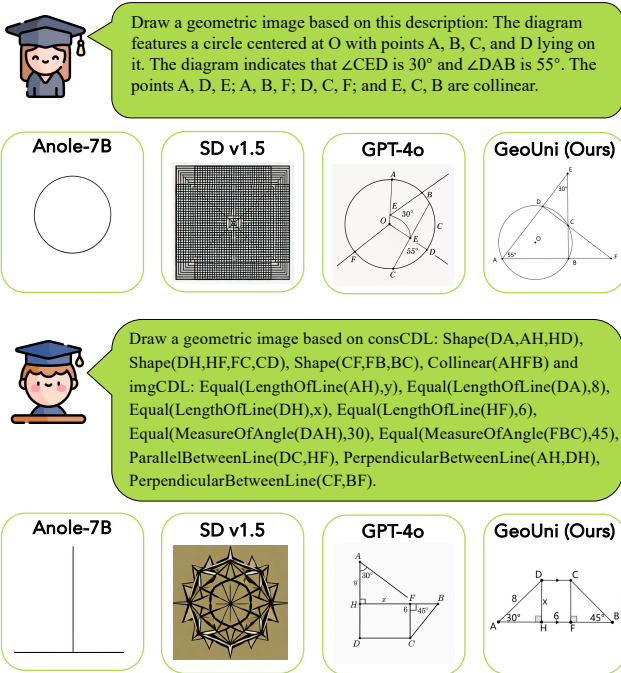


Figure 5: Text-to-Diagram

soft metric. Although some unified and T2I models obtain moderate scores, their generated outputs often fail to resemble valid geometric diagrams in structure or semantics.

As shown in Figure 5, SD v1.5 [29] tends to generate visually rich or stylized content, but its outputs lack geometric structure. Anole-7B [8], fine-tuned from Chameleon-7B [32], is capable of generating simple closed shapes such as circles, but lacks the ability to execute more complex geometry instructions. Additionally, although the recently released GPT-4o [25] demonstrates impressive image generation capabilities, it currently lacks API support. Preliminary results using the web interface show that it can generate visually clear geometric figures, but the outputs do not satisfy precise geometric constraints.

5.5 Reasoning

5.5.1 Metrics. To comprehensively evaluate the geometric reasoning capabilities of the models, we test their performance on both multiple choice and open-ended questions using three public datasets: Formalgeo7K, Geometry3K [21], and GeoQA [6]. For both question types, the models are instructed to reason step-by-step, generate detailed solutions, and present answers in a standardized format to allow accurate comparison with reference solutions. Accuracy metrics are calculated separately based on the question type (multiple choice vs. open-ended) and language (English vs. Chinese), represented as EN-C, CN-C, EN-OE, and CN-OE, respectively.

5.5.2 Results. Table 3 summarizes the reasoning performance across three benchmark datasets in unified models, MLLMs, LLM, and our proposed GeoUni with only a 1.5B-parameter LLM. GeoUni achieves the highest accuracy on English multiple choice questions—75.43%, 71.76%, and 77.99% on Formalgeo7K, Geometry3K, and GeoQA respectively—outperforming significantly larger models such as DeepSeek-V3 [10] and DeepSeek-R1 [9]. While it slightly lags behind DeepSeek-R1 and Qwen2.5-VL-32B [2] in Chinese multiple choice settings, GeoUni remains competitive overall.

On open-ended tasks, GeoUni demonstrates clear advantages in both English and Chinese, particularly through its step-by-step reasoning presented in structured answer formats. For example, on Formalgeo7K (CN-OE), it reaches 55.33%, far surpassing DeepSeek-R1’s 31.71%.

We also observe that many unified models (e.g., Show-o [38], Janus-Pro [7], and Emu3 [36]) perform poorly in both multiple choice and open-ended formats. These models often fail to follow task instructions consistently, leading to performances even worse than random guessing in four-option multiple choice settings. Moreover, models like G-LLaVA-13B [12] do not support Chinese input and thus are only evaluated on English subsets.

5.6 Problem Creation

We further examine GeoUni’s problem-generation capability by prompting it with geometric knowledge points to generate corresponding problems along with appropriate diagrams. As shown in the comparative example between GPT-4o [25] and GeoUni in Figure 6, although GPT-4o can generate images that align with the instructions, the resulting geometric problem is actually unsolvable.

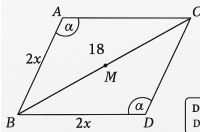
Table 3: Geometric Reasoning Performance Comparison of Various Models Across Different Matrices

Model	Formalgeo7K				Geometry3k				GeoQA			
	EN-C	EN-OE	CN-C	CN-OE	EN-C	EN-OE	CN-C	CN-OE	EN-C	EN-OE	CN-C	CN-OE
Show-o[38]	17.90	2.57	16.57	0.29	16.44	2.31	17.59	0.46	18.93	2.75	15.86	0.16
Chameleon-7B[32]	10.38	5.90	11.81	3.62	9.03	3.94	11.34	3.01	11.33	7.28	12.14	3.01
Janus-Pro-7B[7]	25.62	13.05	8.38	10.57	23.38	12.04	6.25	10.65	27.18	13.75	9.87	10.52
Emu3[36]	45.24	7.14	21.33	3.52	40.05	3.94	21.53	2.08	48.87	9.39	21.20	4.53
SEED-X[14]	13.90	1.33	7.24	3.43	11.81	0.69	5.09	1.85	15.37	1.78	8.74	4.53
GPT-4o[25]	47.52	20.10	49.71	28.48	50.69	23.38	52.08	28.70	45.31	17.80	48.06	28.32
G-LLaVA-13B[12]	47.81	14.10	-	-	5.56	34.95	-	-	56.80	20.06	-	-
Qwen2.5-VL-7B[2]	61.71	26.38	60.48	27.14	64.12	23.84	59.26	22.45	60.03	28.16	61.33	30.42
QWQ-72B[33]	51.62	16.57	20.19	7.14	50.69	23.61	18.52	7.41	52.27	11.65	21.36	6.96
Phi-4-Multimodal[24]	48.19	14.48	43.24	10.00	52.78	10.88	39.81	9.03	44.98	16.99	45.63	10.68
Qwen2.5-Math-1.5B[39]	48.95	24.12	47.43	20.86	45.60	20.14	47.45	21.99	51.29	26.90	47.41	20.06
DS-R1-Distill-Qwen-1.5B[9]	40.38	16.57	38.48	11.71	37.50	19.68	38.19	16.20	42.39	14.40	38.67	8.58
Qwen2.5-Math-72B[39]	70.67	37.90	67.90	37.62	69.44	36.81	65.05	33.33	71.52	38.67	67.90	40.61
DeepSeek-R1[9]	64.86	31.05	77.24	31.71	66.90	36.34	75.23	36.57	63.43	27.35	78.64	28.32
DeepSeek-V3[10]	70.29	33.46	65.71	26.38	71.30	35.66	63.19	28.01	69.58	30.32	67.48	25.24
GeoUni-1.5B(Ours)	75.43	59.81	73.52	55.33	71.76	50.00	69.68	45.27	77.99	66.67	76.21	62.30



Draw a diagram, create a question (multi-choice) and answer based on the given knowledge point.
Knowledge point: definition of a parallelogram, properties of a parallelogram, definition of the midpoint of a line segment, methods for solving equations.

In parallelogram ABCD below, M is the midpoint of line segment BC. What is the value of x?



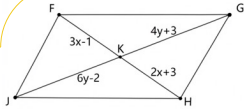
Definition of a Parallelogram
A quadrilateral with both pairs of opposite sides parallel

Properties of a Parallelogram
• Opposite sides are congruent
• Opposite angles are congruent

Definition of the Midpoint of a Line Segment
Divides a line segment into two equal segments

- A) 8
B) 9
C) 12
D) 24

Methods for Solving Equations
Two expressions are equal, set them equal



Question: As shown in the diagram, $FK=3x-1$, $JK=4y+3$, $KH=6y-2$, $KG=2x+3$, $FJHG$ is a $\square$. Find the value of y.
A: 1 B: 2.5 C: 3 D: 5

Reference Answer:
<think> According to the problem statement, FJHG is a parallelogram. Based on the properties of parallelograms, the diagonals bisect each other, which means K is the midpoint of segment JG. Given the length expressions, JK is represented by $4y+3$, and KG is represented by $2x+3$. Since K is the midpoint of JG, we can establish the equation: $4y+3 = 2x+3$. By solving this equation, we find that the value of y is 5/2. </think>
<answer>\boxed{B}</answer>

Figure 6: Problem Creation

In contrast, GeoUni not only generates meaningful geometry problems but also produces accurate geometric diagrams and provides detailed reference answers.

5.7 Ablation Studies

5.7.1 Geo-MAGVIT. We investigate the impact of two key training objectives in Geo-MAGVIT: the topo-perceptual reconstruction loss ($\mathcal{L}_{\text{topo}}$) and the text reconstruction loss ($\mathcal{L}_{\text{text}}$), as shown in Table 4. We evaluate the model’s performance under different configurations by removing either or both losses during training. Removing $\mathcal{L}_{\text{text}}$ significantly impairs the reconstruction of textual elements in the diagram, such as endpoint labels and angle annotations. When both $\mathcal{L}_{\text{topo}}$ and $\mathcal{L}_{\text{text}}$ are removed, the model’s ability to preserve the overall geometric structure degrades notably, as illustrated in Fig. 7.

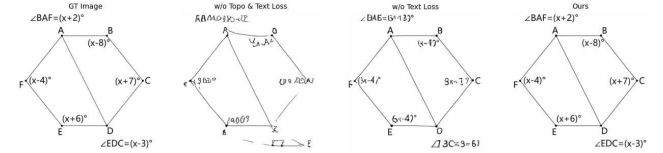


Figure 7: Ablation Study of Geo-MAGVIT

Table 4: Impact of Different Training Loss on Geo-MAGVIT

Topo	Text	C-AA	C-PA	I-AA	I-PA	CI-PA	GPMS
-	-	5.86	1.90	3.28	2.67	0.00	58.66
✓	-	43.24	26.38	18.78	8.00	4.48	71.11
✓	✓	83.10	55.05	75.98	55.71	35.24	91.32

5.7.2 Reasoning. We examine the effects of the stage 3 reasoning enhancement and formalization of diagram information before solving the problem in Formalgeo7K. Table 5 shows that removing the reasoning enhancement (GRPO) in a significant drop in performance across all settings, confirming its importance in boosting model reasoning ability. Likewise, omitting the formalization step also leads to a decrease in both English and Chinese accuracy, especially in open-ended settings.

Table 5: Impact of GRPO and formalization on our model’s performance on the FormalGeo7K

Model	Accuracy			
	EN-C	EN-OE	CN-C	CN-OE
GeoUni (w/o GRPO)	53.81	39.90	55.52	39.33
GeoUni (w/o Formalization)	70.57	53.33	70.19	52.86
GeoUni	75.43	59.81	73.52	55.33

6 Conclusion

In this paper, we propose a unified geometry expert model, **GeoUni**, which integrates geometry problem solving, diagram generation, and problem creation within a single framework. Extensive experimental results demonstrate that GeoUni outperforms existing models in all three tasks. Most importantly, GeoUni makes geometry problem creation a practical reality, bridging the gap between problem solving and teaching. In future works, we aim to explore acceleration techniques to further improve the efficiency of GeoUni, enabling faster geometry problem generation.

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966 [cs.CV] <https://arxiv.org/abs/2308.12966>
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] <https://arxiv.org/abs/2502.13923>
- [3] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. 2024. LoRA Learns Less and Forgets Less. arXiv:2405.09673 [cs.LG] <https://arxiv.org/abs/2405.09673>
- [4] Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. 2024. GeoGPT4V: Towards Geometric Multi-modal Large Language Models with Geometric Image Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 750–766. doi:10.18653/v1/2024.emnlp-main.44
- [5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaoze Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. arXiv:2403.04692 [cs.CV] <https://arxiv.org/abs/2403.04692>
- [6] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 513–523. doi:10.18653/v1/2021.findings-acl.46
- [7] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. arXiv:2501.17811 [cs.AI] <https://arxiv.org/abs/2501.17811>
- [8] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. ANOLE: An Open, Autoregressive, Native Large Multimodal Models for Interleaved Image-Text Generation. arXiv:2407.06135 [cs.CL] <https://arxiv.org/abs/2407.06135>
- [9] DeepSeek-AI, Daya Guo, Dejian Yang, and Others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] <https://arxiv.org/abs/2501.12948>
- [10] DeepSeek-AI, Aixin Liu, Bei Feng, and Others. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL] <https://arxiv.org/abs/2412.19437>
- [11] Hugging Face. 2025. Stability AI / SDXL Turbo. <https://huggingface.co/stabilityai/sdxl-turbo> Accessed: 2025-04-02.
- [12] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjuan Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023. G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv:2312.11370 [cs.CL] <https://arxiv.org/abs/2312.11370>
- [13] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. Making LLaMA SEE and Draw with SEED Tokenizer. arXiv:2310.01218 [cs.CV] <https://arxiv.org/abs/2310.01218>
- [14] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2025. SEED-X: Multimodal Models with Unified Multi-granularity Comprehension and Generation. arXiv:2404.14396 [cs.CV] <https://arxiv.org/abs/2404.14396>
- [15] GeoGebra Team. 2024. GeoGebra. <https://www.geogebra.org/>.
- [16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] <https://arxiv.org/abs/2106.09685>
- [17] Ryan Krueger, Jesse Michael Han, and Daniel Selsam. 2021. Automatically Building Diagrams for Olympiad Geometry Problems. In *CADE*. Springer International Publishing, Cham, 577–588.
- [18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.
- [19] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-IO: A Unified Model for Vision, Language, and Multimodal Tasks. arXiv:2206.08916 [cs.CV] <https://arxiv.org/abs/2206.08916>
- [20] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6774–6786. doi:10.18653/v1/2021.acl-long.528
- [21] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6774–6786. doi:10.18653/v1/2021.acl-long.528
- [22] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. 2025. Open-MAGVIT2: An Open-Source Project Toward Democratizing Autoregressive Visual Generation. arXiv:2409.04410 [cs.CV] <https://arxiv.org/abs/2409.04410>
- [23] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. 2025. UniTok: A Unified Tokenizer for Visual Generation and Understanding. arXiv:2502.20321 [cs.CV] <https://arxiv.org/abs/2502.20321>
- [24] Microsoft. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv:2503.01743 [cs.CL] <https://arxiv.org/abs/2503.01743>
- [25] OpenAI. 2025. Introducing 4.0 Image Generation. <https://openai.com/index/introducing-4o-image-generation/>
- [26] PaddlePaddle. 2025. Introduction to Bayesian Statistics. <https://github.com/PaddlePaddle/PaddleOCR>
- [27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV] <https://arxiv.org/abs/2204.06125>
- [28] Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. 2023. Ocr-vqgan: Taming text-within-image generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. IEEE, Waikoloa, HI, USA, 3689–3698.
- [29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] <https://arxiv.org/abs/2112.10752>
- [30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeek-Math: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] <https://arxiv.org/abs/2402.03300>
- [31] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Emu: Generative Pretraining in Multimodality. arXiv:2307.05222 [cs.CV] <https://arxiv.org/abs/2307.05222>
- [32] Chameleon Team. 2025. Chameleon: Mixed-Modal Early-Fusion Foundation Models. arXiv:2405.09818 [cs.CL] <https://arxiv.org/abs/2405.09818>
- [33] Qwen Team. 2024. QVQ: To See the World with Wisdom. <https://qwenlm.github.io/blog/qvq-72b-preview/>
- [34] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature* 625, 7995 (2024), 476–482.
- [35] Junxiao Wang, Ting Zhang, Heng Yu, Jingdong Wang, and Hua Huang. 2025. MagicGeo: Training-Free Text-Guided Geometric Diagram Generation. arXiv:2502.13855 [cs.CV] <https://arxiv.org/abs/2502.13855>
- [36] Xinlong Wang, Xiaosong Zhang, Zhengxiang Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang,

- and Zhongyuan Wang. 2024. Emu3: Next-Token Prediction is All You Need. arXiv:2409.18869 [cs.CV] <https://arxiv.org/abs/2409.18869>
- [37] Renqiu Xia, Mingsheng Li, Hancheng Ye, Wenjie Wu, Hongbin Zhou, Jiakang Yuan, Tianshuo Peng, Xinyu Cai, Xiangchao Yan, Bin Wang, Conghui He, Botian Shi, Tao Chen, Junchi Yan, and Bo Zhang. 2025. GeoX: Geometric Problem Solving Through Unified Formalized Vision-Language Pre-training. arXiv:2412.11863 [cs.CV] <https://arxiv.org/abs/2412.11863>
- [38] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. arXiv:2408.12528 [cs.CV] <https://arxiv.org/abs/2408.12528>
- [39] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. arXiv:2409.12122 [cs.CL] <https://arxiv.org/abs/2409.12122>
- [40] Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023. A Multi-Modal Neural Geometric Solver with Textual Clauses Parsed from Diagram. arXiv:2302.11097 [cs.AI] <https://arxiv.org/abs/2302.11097>
- [41] Xiaokai Zhang, Na Zhu, Yiming He, Jia Zou, Qike Huang, Xiaoxiao Jin, Yanjun Guo, Chenyang Mao, Yang Li, Zhe Zhu, Dengfeng Yue, Fangzhen Zhu, Yifan Wang, Yiwen Huang, Runan Wang, Cheng Qin, Zhenbing Zeng, Shaorong Xie, Xiangfeng Luo, and Tuo Leng. 2024. FormalGeo: An Extensible Formalized Framework for Olympiad Geometric Problem Solving. arXiv:2310.18021 [cs.AI] <https://arxiv.org/abs/2310.18021>
- [42] Xiaokai Zhang, Na Zhu, Yiming He, Jia Zou, Cheng Qin, Yang Li, and Tuo Leng. 2024. FGeo-SSS: A Search-Based Symbolic Solver for Human-like Automated Geometric Reasoning. *Symmetry* 16, 4 (2024). doi:10.3390/sym16040404
- [43] Zeren Zhang, Jo-Ku Cheng, Jingyang Deng, Lu Tian, Jinwen Ma, Ziran Qin, Xiaokai Zhang, Na Zhu, and Tuo Leng. 2025. Diagram Formalization Enhanced Multi-Modal Geometry Problem Solver. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Hyderabad, India, 1–5. doi:10.1109/ICASSP49660.2025.10889286
- [44] Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. 2025. QLIP: Text-Aligned Visual Tokenization Unifies Auto-Regressive Multimodal Understanding and Generation. arXiv:2502.05178 [cs.CV] <https://arxiv.org/abs/2502.05178>
- [45] Na Zhu, Xiaokai Zhang, Qike Huang, Fangzhen Zhu, Zhenbing Zeng, and Tuo Leng. 2025. FGeo-Parser: Autoformalization and Solution of Plane Geometric Problems. *Symmetry* 17, 1 (2025). doi:10.3390/sym17010008

A Dataset Details

Our GeoUni model is trained on two datasets: FormalGeo7K [41] and SynthGeo228K [43]. The FormalGeo7K dataset contains 5,950 training samples and 1,050 test samples. From the training set, we further separate 1,050 samples specifically for training the Reasoning Enhancement module. SynthGeo228K is a large-scale synthetic dataset comprising geometric diagrams paired with corresponding descriptions. A sample from the FormalGeo7K dataset is shown in Figure 8. Each instance includes problem-text-cn and problem-text-en as the Chinese and English problem descriptions, respectively; consCDL and imgCDL as structured diagram representations; and formalSSS-Solution, which presents a symbolic reasoning process expressed in natural language. Additionally, Figure 9 displays an example from the SynthGeo228K dataset, where each image is accompanied by a formal consCDL representation and a natural language caption.

For diagram tokenization pretraining, we utilize diagrams from the FormalGeo7K training set, which includes 5,950 diagrams, and randomly sample 193,304 diagrams from SynthGeo228K.

Figure 10 illustrates the data augmentation pipeline for multi-task instruction tuning based on the FormalGeo7K dataset. For the Text-to-Diagram (T2D) task, we exploit the permutation invariance of geometric descriptions in the CDLs and the problem text (with questions removed) to perform 10× permutation-based augmentation. We construct both Chinese and English prompt templates to guide diagram generation. Additionally, we employ GPT-4o-mini

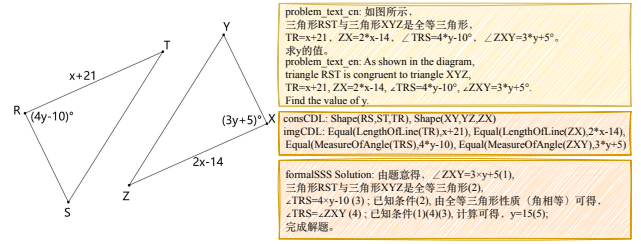


Figure 8: Data Sample of FormalGeo7K.

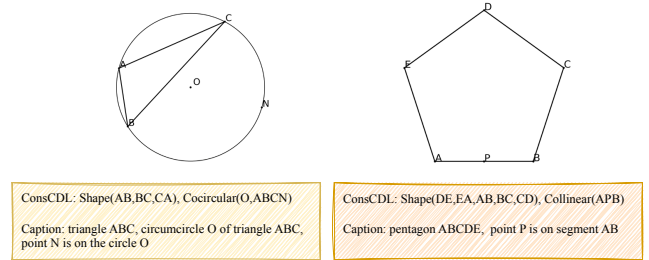


Figure 9: Data Sample of SynthGeo228K.

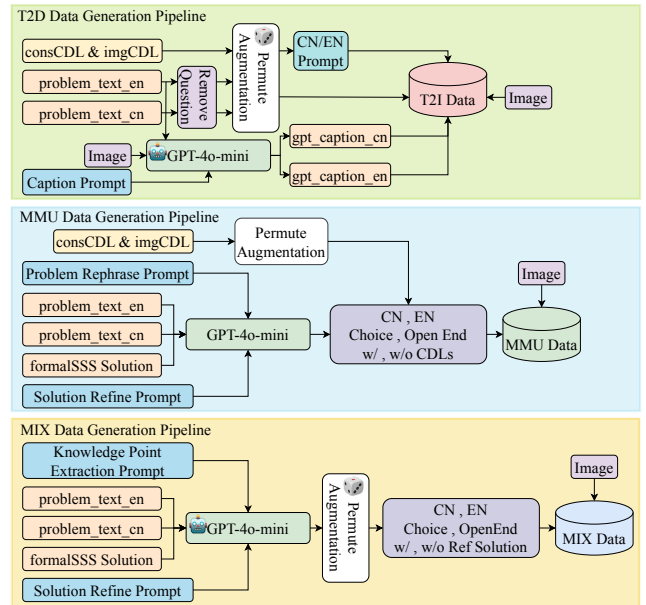


Figure 10: Data Augmentation Pipeline for FormalGeo7K.

to generate diagram descriptions conditioned on both the image content and the associated problem text. This results in 249,900 augmented samples—calculated as $(4 \text{ prompt templates} \times 10 \text{ permutations} + 2 \text{ GPT-generated descriptions}) \times 5,950 \text{ samples}$. An additional 130,900 synthetic samples are incorporated, yielding a total of 380,800 samples for T2D training.

For the geometry problem solving (MMU) task, we define 8 distinct question-answering modes by considering language (Chinese

Table 6: Details of training datasets at different stages.

Stage	Sources	Training Samples
Diagram Tokenization Pretraining	Formalgeo7K (train set) + SynthGeo228K (sampled)	5,950 + 193,204 = 199,254
Multi-Task Instruction Tuning	T2D Formalgeo7K (train set) + SynthGeo228K (sampled)	$(4 \times 10 + 2) \times 5,950 + 130,900 = 380,800$
	MMU Formalgeo7K (train set sampled)	$(4 \times 8) \times 4,900 = 156,800$
	MIX Formalgeo7K (train set sampled)	$(4 \times 8) \times 4,900 = 156,800$
Reasoning Enhancement	Formalgeo7K (train set sampled)	$1,050 \times 8 = 8,400$

or English), question type (multiple-choice or open-ended), and whether geometric diagram formalization is required prior to answering. To enhance the diversity of question phrasing, we apply a 4× rephrasing strategy to each question using GPT-4o-mini. For modes involving pre-formalized geometric diagrams, we further adopt sequence-level augmentation on the corresponding CDL representations. Given that the original formalSSS-solutions are not well-suited for direct model training, we employ GPT-4o-mini to refine them into more human-like solution processes. From the original training set of 5,950 samples, we resample 4,900 instances and construct a dataset of 156,800 MMU samples via the data augmentation pipeline, incorporating all 8 modes and 4 rephrasings per instance.

For the problem augmentation (MIX) task, which involves problem augmentation, we consider eight distinct modes based on language (Chinese or English), question type (multiple-choice or open-ended), and the presence or absence of a reference answer. We employ GPT-4o-mini to extract knowledge points from the original problem text and the corresponding formalSSS solutions. Given the order-invariant nature of these knowledge points, we apply a 4× permutation-based augmentation strategy. Furthermore, similar to the MMU dataset, all reference answers are refined using GPT-4o-mini. In total, our MIX dataset contains $(4 \times 8) \times 4900 = 156,800$ samples after augmentation.

The reasoning enhancement dataset comprises the remaining 1,050 samples from the Formalgeo7K training set in the MMU dataset, which are used for MMU and MIX tasks. By applying various configurations—including question type (multiple-choice or open-ended), language (Chinese or English), and whether formalization is performed beforehand—a total of 8× augmented reinforcement learning samples were constructed, resulting in 8,400 RL training instances.

B Mathematical Proof

The original entropy loss[22] defined as

$$\mathcal{L}_{\text{entropy}}^{\text{old}} = \mathbb{E} [H(f(z_e(T)))] - H(\mathbb{E}[f(z_e(T))])$$

is always non-positive, i.e., $\mathcal{L}_{\text{entropy}}^{\text{old}} \leq 0$, and reaches its minimum value $-\log(C)$, where C denotes the number of codebook entries, and $f(\cdot)$ maps feature representations into a probability distribution over the codebook.

PROOF. Let us denote

$$\mathbf{p}(T) = f(z_e(T)) = (p_1(T), p_2(T), \dots, p_C(T)),$$

representing the probability distribution over the codebook for a given input T . Then the entropy loss becomes:

$$\mathcal{L}_{\text{entropy}}^{\text{old}} = \mathbb{E}_T [H(\mathbf{p}(T))] - H(\mathbb{E}_T [\mathbf{p}(T)]). \quad (20)$$

Note that for $p \in (0, 1)$, the function $g(p) = -p \log(p)$ has a second derivative $g''(p) = -\frac{1}{p} < 0$, implying that $g(p)$ is convex. Therefore, the entropy function

$$H(\mathbf{p}) = \sum_{i=1}^C -p_i \log(p_i) \quad (21)$$

$$= \sum_{i=1}^C g(p_i) \quad (22)$$

is a sum of convex functions, and hence also convex. By Jensen’s inequality, we obtain:

$$\mathbb{E}_T [H(\mathbf{p}(T))] \leq H(\mathbb{E}_T [\mathbf{p}(T)]), \quad (23)$$

which leads to $\mathcal{L}_{\text{entropy}}^{\text{old}} \leq 0$, indicating a contradiction with the common expectation that loss functions are non-negative.

Furthermore, using the property of entropy that $0 \leq H(\mathbf{p}) \leq \log(C)$, we derive:

$$\mathbb{E}_T [H(\mathbf{p}(T))] \geq 0, \quad (24)$$

$$-H(\mathbb{E}_T [\mathbf{p}(T)]) \geq -\log(C). \quad (25)$$

Summing the two inequalities yields:

$$\mathbb{E}_T [H(\mathbf{p}(T))] - H(\mathbb{E}_T [\mathbf{p}(T)]) \geq -\log(C), \quad (26)$$

i.e., $\mathcal{L}_{\text{entropy}}^{\text{old}} \geq -\log(C)$.

To demonstrate that the lower bound is tight, we consider a specific case. Suppose we have $B = C$ samples, and the sample mean is used to approximate the expectation. In this case, the loss becomes:

$$\mathcal{L}_{\text{entropy}}^{\text{old},*} = \frac{1}{B} \sum_{k=1}^B H(\mathbf{p}(T_k)) - H\left(\frac{1}{B} \sum_{k=1}^B \mathbf{p}(T_k)\right). \quad (27)$$

For $k = 1, \dots, C$, let us define

$$\mathbf{p}(T_k) = (0, \dots, 0, 1, 0, \dots, 0), \quad (28)$$

where the 1 appears in the k -th position, and the rest are zeros. In this case, we have:

$$H(\mathbf{p}(T_k)) = 0, \quad (29)$$

$$H\left(\frac{1}{B} \sum_{k=1}^B \mathbf{p}(T_k)\right) = H\left(\frac{1}{C}, \dots, \frac{1}{C}\right) \quad (30)$$

$$= \log(C). \quad (31)$$

Therefore,

$$\mathcal{L}_{\text{entropy}}^{\text{old},*} = \frac{1}{B} \sum_{k=1}^B H(\mathbf{p}(T_k)) - H\left(\frac{1}{B} \sum_{k=1}^B \mathbf{p}(T_k)\right) \quad (32)$$

$$= 0 - \log(C) \quad (33)$$

$$= -\log(C). \quad (34)$$

This concludes the proof that the minimum of $\mathcal{L}_{\text{entropy}}^{\text{old}}$ is $-\log(C)$. $\square$

C Examples for Diagram Reconstruction

We demonstrate the superior performance of Geo-MAGVIT in diagram reconstruction, as shown in Figure 11.

D Examples for Text to Diagram

We demonstrate the superior performance of GeoUni in Text to Diagram based on different prompt formats, as shown in Figures. 12, 13, 14, 15, 16, and 17.

E Examples for Problem Creation

We demonstrate the superior performance of GeoUni in Problem Creation based on English and Chinese prompts, compared to GPT-4o, as shown in Figures. 18, 19, and 20.

Original			
UniTok			
QLIP			
MAGVIT			
Geo-MAGVIT (Ours)			

Figure 11: Diagram Reconstruction

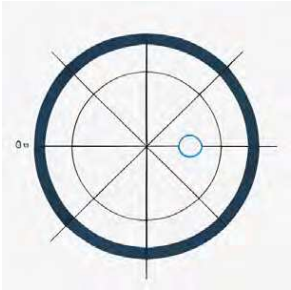
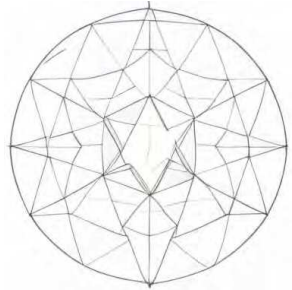
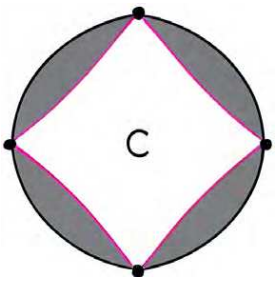
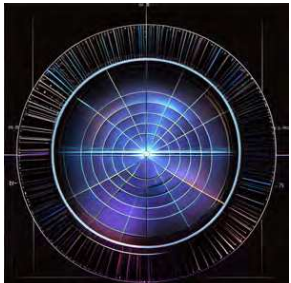
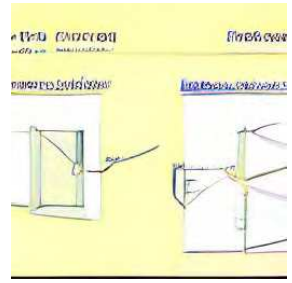
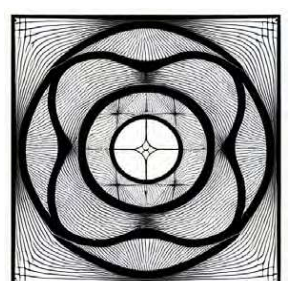
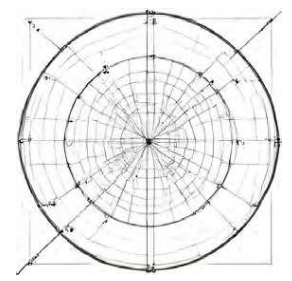
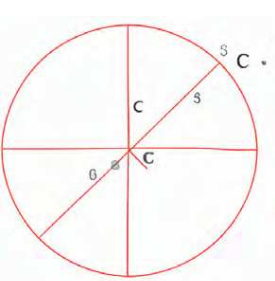
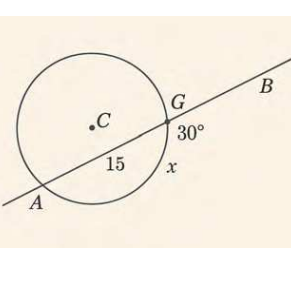
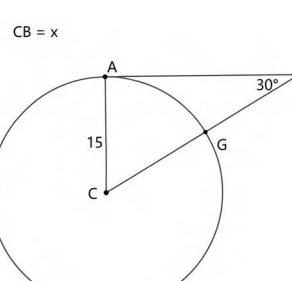
Draw a geometric image based on this description: $CA=15$, $CB=x$, $\angle GBA=30^\circ$, C is the center of circle C, and the tangent to circle C is BA.	Show-O 	JanusPro-7B 
Anole-7B 	Emu3 	UNIFIED-IO 
PixArt-Σ 	SD v1.5 	SDXL Turbo 
DALL-E 2 	GPT4o 	GeoUni (Ours) 

Figure 12: Text-To-Diagram-Showcase-1

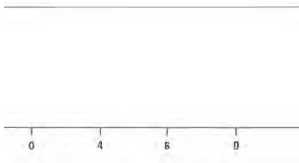
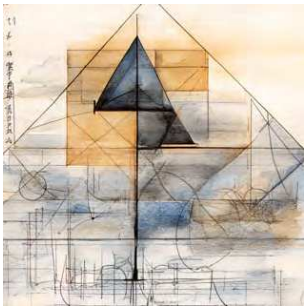
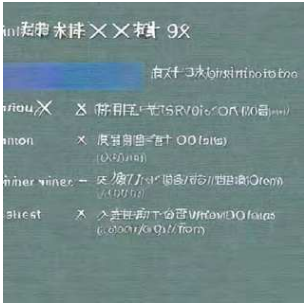
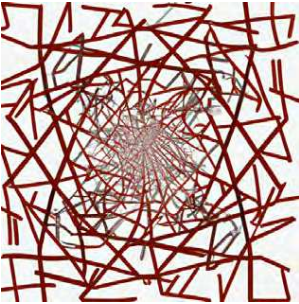
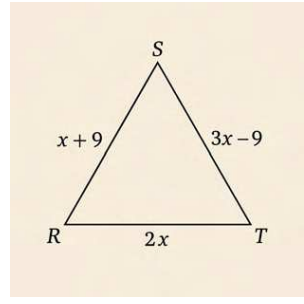
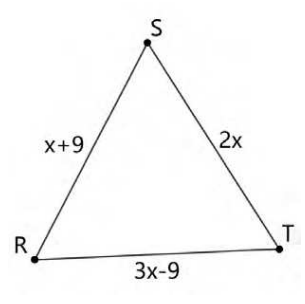
根据这段描述绘制一张几何图像： $RS=x+9$, $RT=3x-9$, $ST=2x$, $\triangle SRT$ 为等边$\triangle$.	Show-O 	JanusPro-7B 
Anole-7B Stor stre the loda+renty,R:SO -no lv1 In arago lmeratig rfrle 	Emu3 	UNIFIED-IO 
PixArt-Σ 	SD v1.5 	SDXL Turbo 
DALL-E 2 	GPT4o 	GeoUni (Ours) 

Figure 13: Text-To-Diagram-Showcase-2

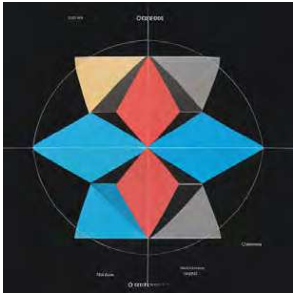
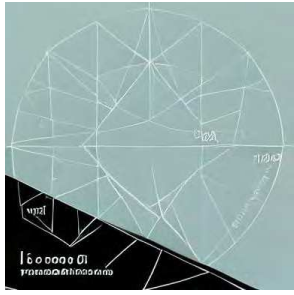
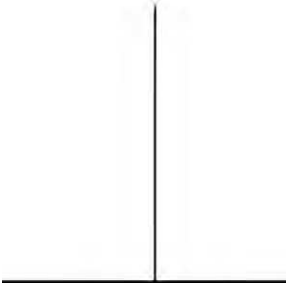
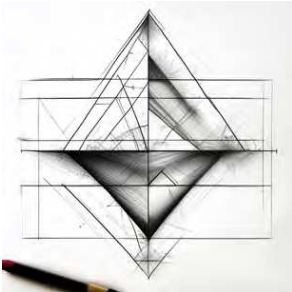
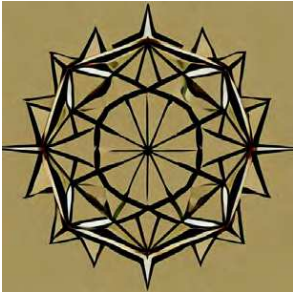
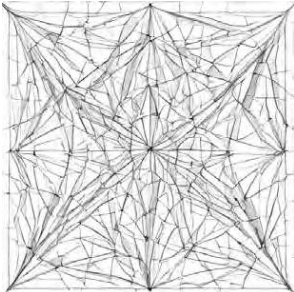
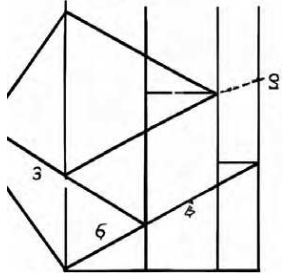
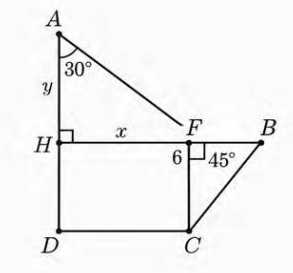
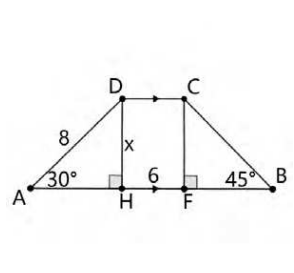
Draw a geometric image based on consCDL: Shape(DA,AH,HD), Shape(DH,HF,FC,CD), Shape(CF,FB,BC), Collinear(AHFB) and imgCDL: Equal(LengthOfLine(AH),y), Equal(LengthOfLine(DA),8), Equal(LengthOfLine(DH),x), Equal(LengthOfLine(HF),6), Equal(MeasureOfAngle(DAH),30), Equal(MeasureOfAngle(FBC),45), ParallelBetweenLine(DC,HF), PerpendicularBetweenLine(AH,DH), PerpendicularBetweenLine(CF,BF).	Show-O 	JanusPro-7B 
Anole-7B 	Emu3 	UNIFIED-IO 
PixArt-Σ 	SD v1.5 	SDXL Turbo 
DALL-E 2 	GPT4o 	GeoUni(Ours) 

Figure 14: Text-To-Diagram-Showcase-3

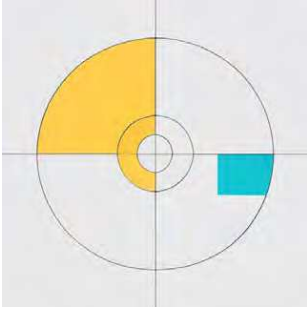
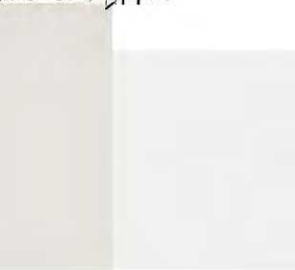
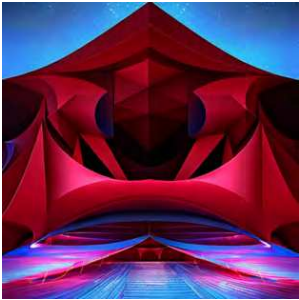
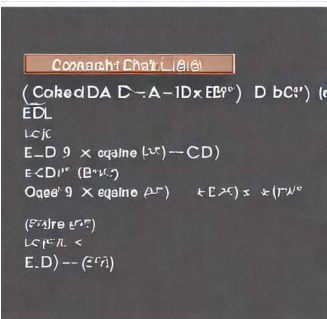
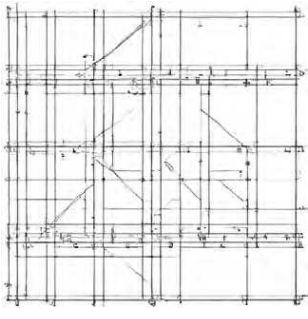
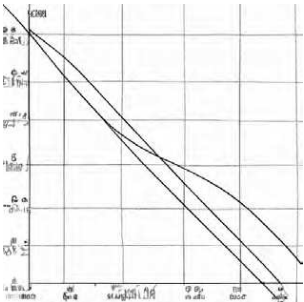
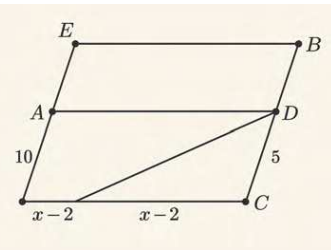
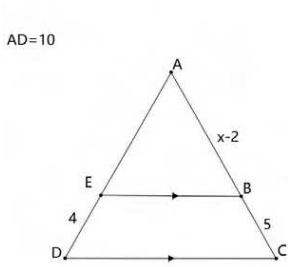
根据 consCDL: Shape(AE,EB,BA), Shape(ED,DC,CB,BE), Collinear(AED), Collinear(ABC) 和 imgCDL: Equal(LengthOfLine(AD),10), Equal(LengthOfLine(BA),x-2), Equal(LengthOfLine(BC),5), Equal(LengthOfLine(ED),4), ParallelBetweenLine(EB,DC) 绘制一张几何图像。	Show-O 	JanusPro-7B 
Anole-7B 	Emu3 	UNIFIED-IO 
PixArt-Σ 	SD v1.5 	SDXL Turbo 
DALL-E 2 	GPT4o 	GeoUni(Our) 

Figure 15: Text-To-Diagram-Showcase-4

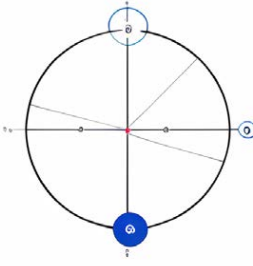
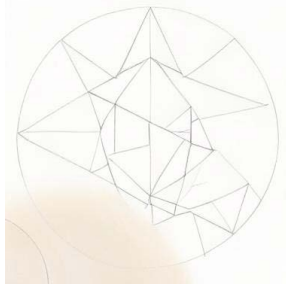
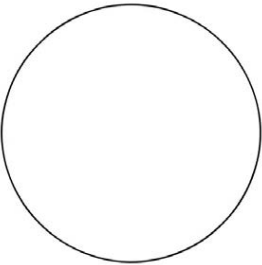
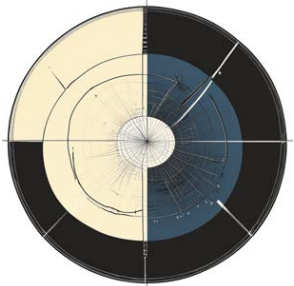
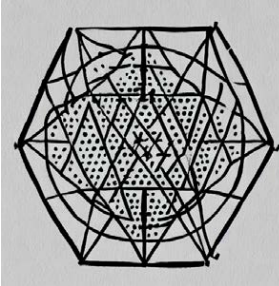
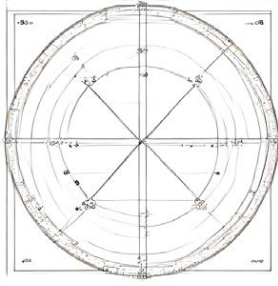
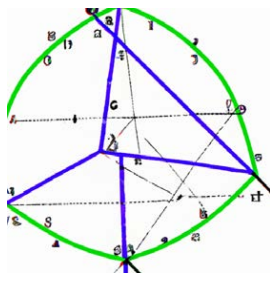
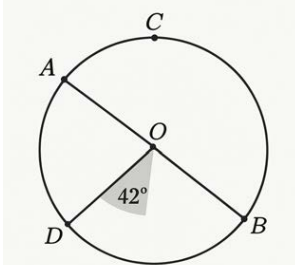
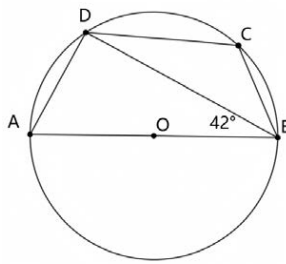
Draw a geometric image based on this description: The diagram features a circle centered at O, with points A, B, C, and D lying on the circle. It includes a diameter AB and specifies $\angle ABD = 42^\circ$.	Show-O 	JanusPro-7B 
Anole-7B 	Emu3 	UNIFIED-IO 
PixArt-Σ 	SD v1.5 	SDXL Turbo 
DALL-E 2 	GPT4o 	GeoUni(Ours) 

Figure 16: Text-To-Diagram-Showcase-5

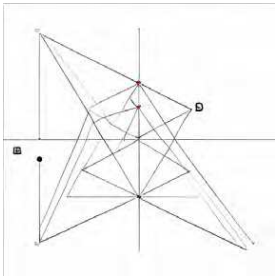
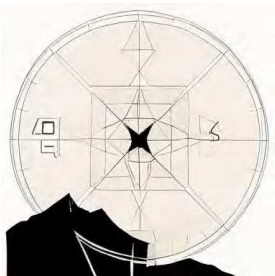
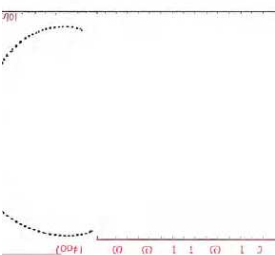
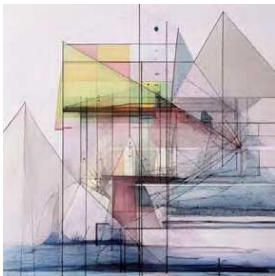
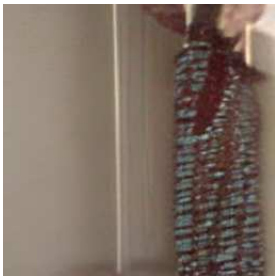
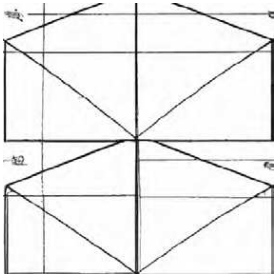
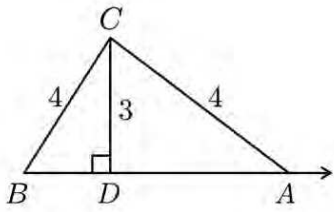
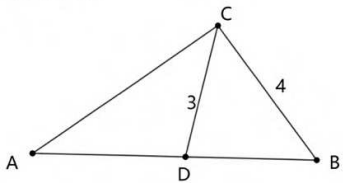
根据这段描述绘制一张几何图像： 该图涉及一组共线点 A、D 和 B，形成一个几何图形的一个部分，其中角 BCA 为直角。线段 BC 和 CD 的长度分别为 4 和 3 个单位，线段 CD 在三角形 CAB 中充当中线，作为该配置的一部分。	Show-O 	JanusPro-7B 
Anole-7B 	Emu3 	UNIFIED-IO 
PixArt-Σ 	SD v1.5 	SDXL Turbo 
DALL-E 2 	GPT4o 	GeoUni(Ours) $\angle BCA = 90^\circ$ 

Figure 17: Text-To-Diagram-Showcase-6

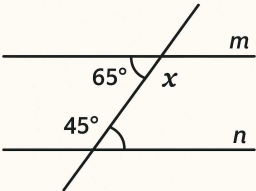
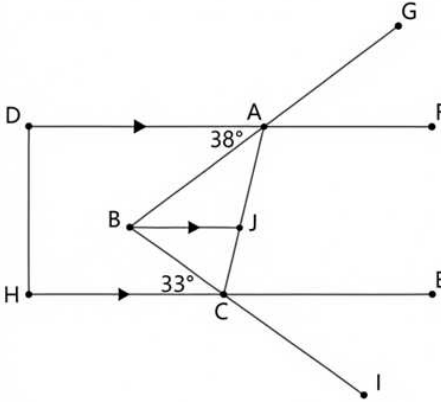
Prompt	根据提供的知识点绘图、出题（选择题）并解答。知识点：平行线的性质，角的定义，角的相加性质，角度计算。
GPT-4o	 Lines m and n are parallel. What is the value of x? A. 45° B. 65° C. 110° D. 115° Answer: C. 110°
GeoUni	 题目：如图所示，$\angle BAD=38^\circ$，$\angle HCB=33^\circ$，$BJ\parallel HC$，DA 平行于 BJ。求 $\angle ABC$ 的大小。 A: 33 B: 38 C: 61 D: 71 参考答案： <think> 根据题意，已知 DA 平行于 BJ，$\angle BAD=38^\circ$，BJ 平行于 HC，$\angle HCB=33^\circ$。 首先，由于 DA 平行于 BJ，根据平行线的性质，我们可以得出 $\angle ABJ=\angle BAD$。因此，$\angle ABJ=38^\circ$。接下来，由于 BJ 平行于 HC，同样适用平行线的性质，我们得到 $\angle HCB=\angle JBC$。所以，$\angle JBC=33^\circ$。最后，根据角的相加法则，我们可以得到 $\angle ABC=\angle ABJ+\angle JBC$。代入已知的角度，$\angle ABC=38^\circ+33^\circ=71^\circ$。因此，最终结果为 $\angle ABC=71^\circ$。</think> <answer>\boxed{D}</answer>

Figure 18: Problem-Creation-Showcase-1

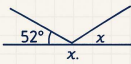
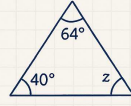
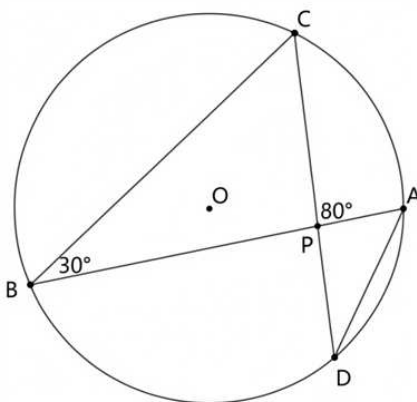
Prompt	根据提供的知识点绘图、出题并解答。知识点：邻补角定理、圆周角定理、三角形内角和定理。
GPT-4o	Using the Adjacent Supplementary Angle Theorem, Inscribed Angle Theorem, Triangle Sum Theorem Linear Pair Theorem  Find x. By the Linear Pair Theorem, Find y. $x + 52^\circ = 180^\circ$ $x = 180^\circ - 52^\circ = 128^\circ$ By the Inscribed Angle Theorem, $y = 128^\circ / 2 = 64^\circ$ Triangle Sum Theorem  Find z. By the Triangle Sum Theorem, $64^\circ + 40^\circ + z = 180^\circ$ $z = 180^\circ - 104^\circ = 76^\circ$
GeoUni	 题目：如图所示，$\angle CBP=30^\circ$，$\angle CPA=80^\circ$，O 是 $\odot O$ 的圆心。求 $\angle DAB$ 的大小。 参考答案： <think> 根据题意，我们知道 $\angle CBP=30^\circ$，$\angle CPA=80^\circ$。根据弧的性质，同弧所对的圆周角等于圆心角的一半，得到 $\angle CBP=\angle OAC/2$。因此，可以计算出 $\angle OAC=60^\circ$。同样地，根据弧的性质，我们可以得到 $\angle PDA=\angle OAC/2$，这样就能得出 $\angle PDA=30^\circ$。接下来，利用邻补角定理，我们有 $\angle APD=180^\circ-\angle CPA$，因此可以得出 $\angle APD=100^\circ$。最后，利用三角形内角和为 180° 的性质，我们可以写出 $\angle APD=\angle DAB+\angle PDA$，通过已知的角度计算得出 $\angle DAB=70^\circ$。解题过程顺利完成。</think> <answer>\boxed{70}</answer>

Figure 19: Problem-Creation-Showcase-2

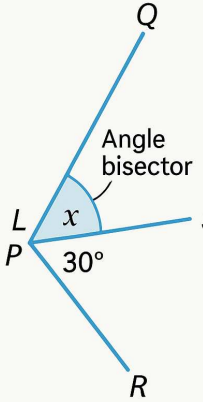
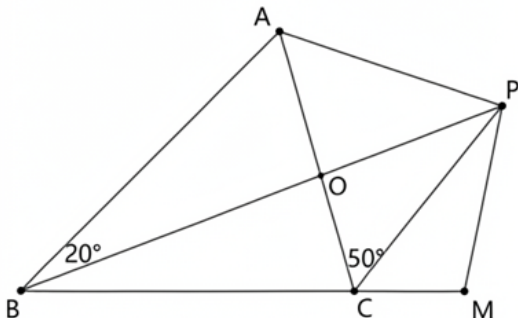
Prompt	Draw a diagram, create a question and answer based on the given knowledge point. Knowledge point: definition of angle bisector, theorem of adjacent supplementary angles, properties of angles, theorem of the sum of interior angles in a triangle, properties of complementary angles, properties of angle equality, operations of sum and difference of angles.	
GPT-4o	If $\angle PQR = 30^\circ$, find the value of x.  • Adjacent supplementary angles theorem • Properties of angles • Angle sum theorem of a triangle • Properties of complementary angles • Properties of angle equality • Operations of sum and difference of angles	
GeoUni	 Question: As shown in the diagram, $\angle ABP=20^\circ$, $\angle ACP=50^\circ$, BP bisects $\angle ABC$, CP bisects $\angle ACM$. Find the measure of $\angle CAB$. Reference Answer: <think> According to the problem statement, BP is the angle bisector of $\angle ABC$, and $\angle ABP=20^\circ$. Therefore, we can deduce that $\angle PBC=20^\circ$. Based on the properties of angles, we can express $\angle ABC$ as $\angle ABP$ plus $\angle PBC$, leading to $\angle ABC=40^\circ$. Next, CP is the angle bisector of $\angle ACM$, and $\angle ACP=50^\circ$, allowing us to conclude that $\angle PCM=50^\circ$. Using the properties of angles, we can express $\angle ACM$ as $\angle ACP$ plus $\angle PCM$, resulting in $\angle ACM=100^\circ$. According to the supplementary angle theorem, we can express $\angle ACM$ as $-\angle BCA$ plus 180°, which leads to $\angle BCA=80^\circ$. Finally, utilizing the property of the sum of angles in a triangle, we express $\angle ABC$ as $-\angle BCA - \angle CAB + 180^\circ$, substituting in known values to find $\angle CAB=60^\circ$. Thus, the entire solution process is complete.</think> <answer>\boxed{60}</answer>	

Figure 20: Problem-Creation-Showcase-3

Prompt	Draw a diagram, create a question and answer based on the given knowledge point. Knowledge point: the relationship between central angles and inscribed angles, the properties of tangents to a circle, the theorem of the sum of interior angles of a triangle.
GPT-4o	• the relationship between central angles and inscribed angles • the properties of tangents to a circle • The theorem of the sum of the interior angles of a triangle In the figure, what is the value of x? Answer: 55°
GeoUni	Question: As shown in the diagram, $\angle CDB = 20^\circ$, O is the center of $\odot O$, EC is the tangent to circle O. Find the measure of $\angle BEC$. Reference Answer: <think> According to the problem statement, we first know that $\angle CDB = 20^\circ$. Let O be the center of circle O, and EC be the tangent to circle O. By the properties of circles, we know that $\angle COB$ equals the measure of arc BC. Based on the property of arcs, the angle subtended by the same arc at the circumference equals half the angle at the center; thus, we have $\angle CDB = \angle COB / 2$. Given that $\angle CDB = 20^\circ$, we can conclude that $\angle COB = 40^\circ$. Next, since $\angle COB = \angle COB$, we have $\angle COB = 40^\circ$. Considering the properties of triangles, the sum of the internal angles is 180°, leading to the equation $\angle BEC = 180^\circ - \angle COB - \angle ECO$. Since EC is a tangent, and according to the properties of tangents, we know that EC is perpendicular to OC, hence $\angle ECO = 90^\circ$. Combining the previous results, we can calculate $\angle BEC = 50^\circ$. The solution process is complete.</think> <answer>\boxed{50}</answer>

Figure 21: Problem-Creation-Showcase-4