

Concise Geometric Description as a Bridge: Unleashing the Potential of LLM for Plane Geometry Problem Solving

Supplementary Material

Table 8. **Effect of rollout number N in the GRPO stage.** We perform the experiments on Qwen2.5-VL 7B. Accounting for both training cost and performance, we set $N = 8$.

Performance		$N = 5$	$N = 8$	$N = 10$
TextCDL	Recall	98.5	99.1	98.3
	Precision	98.4	99.1	98.3
ImgCDL	Recall	96.9	97.0	97.0
	Precision	96.9	96.9	97.0
ConsCDL	Recall	89.1	92.7	91.0
	Precision	87.9	92.1	90.3
Formalgeo-Rec-CoT		80.9	83.2	82.5

Table 9. **Effect of Reward Weights α, γ in GRPO.** We perform the experiments on Qwen3-VL 8B. According to results, we set $\alpha = 0.1, \gamma = 0.5$.

α	γ	TextCDL		ImgCDL		ConsCDL		Acc.
		Re.	Pre.	Re.	Pre.	Re.	Pre.	
0.2	0.4	98.3	98.9	96.8	96.8	96.1	96.0	84.4
0.4	0.3	98.4	99.0	97.2	97.0	96.1	96.0	83.6
0.1	0.5	99.2	99.3	97.0	96.9	95.9	95.8	85.7

6. More Ablations

Effect of rollout number N in GRPO. In order to validate the effect of rollout number N in the GRPO stage, we perform an ablation study on Qwen2.5-VL 7B in Table 8. Setting $N = 10$ yields no performance gain on CDL generation and slightly degrades the problem solving accuracy. Moreover, it brings an extra 80 hours of training time compared with $N = 8$. Considering the training cost and the performance, we set $N = 8$.

Effect of Reward Weights α, γ in GRPO. In order to validate the effect of reward weights α, γ in the GRPO stage, we perform an ablation study on Qwen3-VL 8B in Table 9. Considering the CDL generation performance and the solving accuracy, we set $\alpha = 0.1, \gamma = 0.5$.

Effect of learning rate in SFT and RL stages. During the training process, we adopt the learning rate of $1.0 \times e^{-5}$ for

Table 10. **Superior Performance on Original Formalgeo7k v2.** We further conduct both training and evaluation on the original Formalgeo7k v2. Compared with previous methods, our superior performance on the original Formalgeo7k v2 validates the effectiveness of our framework.

Methods	TextCDL		ImgCDL		ConsCDL	
	Re.	Pre.	Re.	Pre.	Re.	Pre.
FgeoParser (Diag.)	-	-	77.5	-	87.0	-
FgeoParser (Text)	96.5	-	-	-	-	-
Diagram Formalizer	-	-	92.9	-	90.3	-
Ours (Formalgeo7k)	98.2	98.3	97.0	97.0	91.0	90.3
Ours (Rec-CoT)	99.1	99.1	97.0	96.9	92.7	92.1

SFT and $1.0 \times e^{-6}$ for RL. We further conduct ablations on various choices of learning rate. Results confirm that the chosen hyperparameters yield the best results, and the performance remains relatively stable across various choices.

Superior Performance on Original Formalgeo7k v2. In Table 10, we conduct both training and evaluation on the original Formalgeo7k v2 (“Ours (Formalgeo7k)” in the table). The results demonstrate that our method outperforms previous CDL generation approaches even when using the original data, validating the effectiveness of our framework. Furthermore, training with the refined data yields better performance, confirming that high-quality CDL annotations are also beneficial.

Empirical analysis of how concise descriptions benefit MLLM’s learning. We further compare the performance of concise CDL against the expanded version on the Formalgeo-Rec-CoT training set. While results in Table 11 show that the expanded version achieves slightly higher CDL matching scores on the training set, its performance on the validation set is inferior to our concise CDL (as shown in Table 5). This phenomenon shows the relatively poor generalization of the expanded version compared to ours. The expanded version introduces unnecessary or redundant information, which unexpectedly enlarges the search space. Consequently, with relatively limited samples, MLLM tends to memorize training samples rather than learning generalizable patterns. In contrast, the search space

Table 11. **Concise descriptions (ours) v.s. Expanded version on the training set.** While the expanded version achieves slightly higher scores than the concise CDL on the training set, its performance on the validation set (shown in Table 5) degrades, which shows poor generalization.

Type	TextCDL		ImgCDL		ConsCDL	
	Re.	Pre.	Re.	Pre.	Re.	Pre.
Concise	99.3	99.4	98.7	98.5	93.8	94.1
Expanded	99.8	99.8	98.7	98.7	94.0	94.1

of concise CDL is smaller, which eases the generalization. **Generalization of CDL for various LLM Solvers.** In order to validate the generalization capability of our proposed diagram, we further perform an ablation for various LLM Solvers. Taking the generated CDL from our MLLM Interpreter (Qwen2.5-VL 7B) as inputs, we utilize various LLMs to serve as solvers, including Qwen3 30B, Qwen3 32B, DeepSeek-V3.1-Terminus, and GLM-4.6. Results in Table 12 demonstrate that our proposed paradigm achieves overall superior performance against previous methods across LLM solvers, validating its strong generalization. Qwen3 30B yields the best result and is therefore selected as the LLM solver in our main paper.

Solving Accuracy Comparison with Diagram Formalizer on the same LLM Solver. We further evaluate end-to-end solving accuracy improvements against DFE-GPS’s Diagram Formalizer [42] across both in-domain (Formalgeo7k v2) and Out-Of-Domain (Unigeo & Mathvista) benchmarks. CDL generation results of Diagram Formalizer are obtained by running its official code. In Table 13, using the **same LLM solver (Qwen3 30B)**, ours yields consistent improvement across all benchmarks, further validating our generalization capability and confirming that enhancement on CDL generation indeed leads to better solving performance.

7. Proof for CDL’s Conciseness

In this section, we provide a proof to demonstrate the conciseness of Conditional Declaration Language (CDL) compared with general textual descriptions.

Generally, a textual description of a geometric input can be decomposed into three components: 1) shape descriptions that depict geometric shapes, *e.g.*, line segments, angles, triangles, *etc.*, 2) relation descriptions that reflect the positional and algebraic relationships between shape elements, and quantitative metric (*e.g.*, the line segment length and the angle degree), and 3) irrelevant words that are irrelevant to any geometric shapes or relations, *e.g.*, “a”, “the”, *etc.*

Table 12. **Generalization of CDL for various LLM Solvers.** We perform experiments with various LLM Solvers on Formalgeo-Rec-CoT. All LLM Solvers take the generated CDL from our MLLM Interpreter (Qwen2.5-VL 7B) as inputs to perform reasoning.

Models		Formalgeo
Closed-Source MLLMs		
GPT-4o		58.0
Claude-Sonnet-4		69.1
Claude-Opus-4.1		69.1
Gemini2.5-Flash		80.5
Gemini2.5-Pro		81.8
Open-Source MLLMs		
Qwen2.5-VL 32B		57.3
Qwen3-VL 30B		80.1
GLM4.1-V		73.4
GeoUni		59.8
DFE-GPS		75.3
Ours		
MLLM Interpreter	+Qwen3 32B	82.5
	+GLM-4.6	82.3
	+DeepSeek V3.1	79.4
	+Qwen3 30B	83.2

Table 13. **Solving Accuracy Comparison with Diagram Formalizer on the same LLM Solver.** We evaluate end-to-end solving accuracy improvements against DFE-GPS’s Diagram Formalizer on the same LLM Solver, Qwen3 30B. Results demonstrate the generalization capability of our method and confirm that enhancement on CDL generation indeed leads to better solving performance.

Models	LLM	Formalgeo	Unigeo	MathV.
DFE-GPS	Qwen3	80.3	79.9	69.0
Ours	Qwen3	85.7	84.0	80.8

For a specific geometric input, let H , R , and O denote the sets of all possible shapes, relations, and irrelevant words, respectively, under specific constraints or rules (*e.g.*, CDL rules).

(1) **Shape Descriptions:** In CDL, the shape description

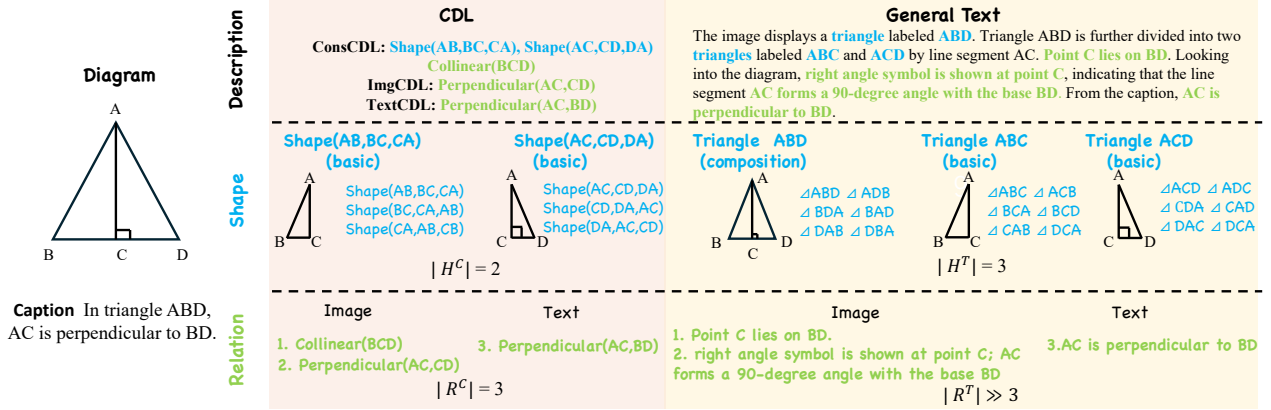


Figure 4. **Illustration of CDL’s Conciseness.** A textual description of a geometric diagram can be decomposed into **shape descriptions** (H), **relation descriptions** (R), and **irrelevant words** (O). We use “basic” to denote the triangles that cannot be decomposed further, and “composition” to denote the complex triangles composed of two basics. Under specific constraints in the CDL rule, $|H^C|$, $|R^C|$, and $|O^C|$ are minimal to $|H^T|$, $|R^T|$, and $|O^T|$, respectively. Therefore, for any given geometric diagram, a CDL description is significantly more concise than a general textual description.

set H^C consists of closed shapes, angles, and line segments. There are two constraints: a) the closed shape cannot be decomposed further into other closed shapes; b) the angle does not exist in any closed shape in the diagram; the line segment does not exist in any angle or closed shape in the diagram. While in general textural descriptions H^T , there is no such constraint. Thus, in H^T , in addition to elements in H^C , there may also exist closed shapes that can be decomposed further and angles or line segments that exist in a specific closed shape. Consequently, $H^C \subseteq H^T$ and $|H^C| \leq |H^T|$.

(2) **Relation Descriptions:** Suppose we need to depict K relations to solve our problem. For each piece of relation, there is only one possible description for such a relation in CDL. So in total, $|R^C|$ of CDL equals the number of relations K . However, there are many choices for describing the same relation in general textual descriptions. Thus, in total $|R^T| \gg K$.

(3) **Irrelevant Words:** As CDL is a description language with a predefined format, CDL doesn’t contain any irrelevant words. Thus, the O^C for CDL and O^T for general text satisfy $|O^C| = 0 \ll |O^T|$.

Therefore, for any given geometric input, the following inequality satisfies:

$$|CDL| = |H^C| + |R^C| + |O^C| \ll |H^T| + |R^T| + |O^T| = |Text|.$$

This inequality proves that for a given geometric input, the number of all possible descriptions in CDL $|CDL|$ is much smaller than that in general textual descriptions $|Text|$, *i.e.*, the CDL description is more concise than the general textual description. We provide an illustration to demonstrate this in Fig. 4.

8. More Qualitative Results

In this section, we provide examples of various benchmarks, including Formalgeo-Rec-CoT, Unigeo, and Mathvista.

