Thinking with Constructions: A Benchmark and Policy Optimization for Visual-Text Interleaved Geometric Reasoning

Anonymous ACL submission

Abstract

Geometric reasoning inherently requires **"thinking with constructions"**—the dynamic manipulation of visual aids to bridge the gap between problem conditions and solutions. However, existing Multimodal Large Language Models (MLLMs) are largely confined to passive inference with static diagrams, lacking the strategic knowledge of when and how to construct effective visual aids. To address this, we present a framework for **Visual-Text Interleaved Chain-of-Thought**. We first introduce **GeoAux-Bench**, the first benchmark comprising 4,334 geometry problems that aligns textual construction steps with ground-truth visual updates. Our pilot study reveals two critical insights: (1) interleaved visual-textual aids outperform single-modality counterparts, which cannot losslessly capture geometric synergy; and (2) valid constructions act as entropy reducers, strongly correlating with reduced reasoning perplexity. Building on these findings, we propose **Action Applicability Policy Optimization (A^{2}PO)**, a reinforcement learning paradigm for mastering strategic construction. A employs Adaptive Reward Shaping to regulate the timing and quality of visual aids via counterfactual sampling to distinguish necessary from redundant constructions. Experiments demonstrate our approach enables MLLMs to leverage selective auxiliary constructions, yielding a 3.51% gain over strong baselines. Code and data are available on GitHub [1].

1 Introduction

Recent advancements in Large Language Models (LLMs) have demonstrated remarkable proficiency in mathematical reasoning (Shao et al., 2024; Yang et al., 2024; Chen et al., 2025a), largely driven by the Chain-of-Thought (CoT) technique (Wei et al., 2022). However, geometry problem solving remains a significant hurdle. Unlike algebraic tasks that rely on symbolic manipulation, geometric reasoning is intrinsically multimodal (Lu et al., 2021; Kazemi et al., 2023): human experts do not merely read static diagrams; they solve problems by constructing and manipulating visual aids (e.g., drawing auxiliary lines) to bridge the gap between conditions and solutions. This process of **auxiliary construction** is the quintessential embodiment of *"thinking with images,"* where the visual context dynamically evolves to reveal hidden geometric properties (Chern et al., 2025; Li et al., 2025b).

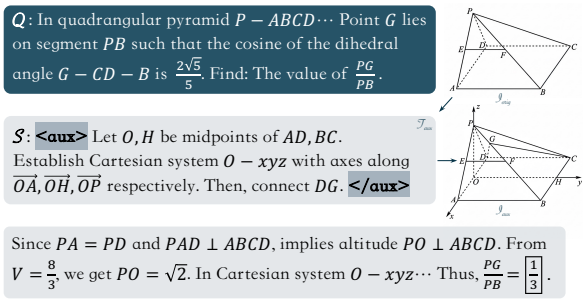


Figure 1: A representative sample from GeoAux-Bench. The solution trajectory is structured in a **visual-text interleaved format**: the textual auxiliary construction step (e.g., <aux>. . . </aux>) is explicitly paired with a corresponding auxiliary diagram (I_{aux}).

To emulate this process, the community has actively explored **Visual Chain-of-Thought (VCoT)**. However, existing paradigms generally fall into three categories, each facing distinct limitations: (1) **Agent-based approaches** manipulate geometric code to render diagrams but often rely on ground-truth code inputs, diverging from the natural visual perception of raw images (Hu et al., 2024; Wang et al., 2025b); (2) **Formal Abstraction** methods convert diagrams into formal languages, acting as a *lossy compression* that strips away visual intuition and risks hallucination (Sharma et al., 2024; Trinh et al., 2024; Yang et al., 2025); and (3) **Unified Multimodal Models** attempt to natively generate visual thoughts (Shi et al., 2025; Li et al.,

[1] https://anonymous.4open.science/r/GeoAux-5863

2025a). While promising, even SOTA models like Nano-Banana-Pro (Team et al., 2025) suffer from pixel-level structural hallucinations that mislead the reasoning process. Consequently, most MLLMs are confined to a passive static inference mode, unable to update their visual context to match their reasoning steps.

This disconnect highlights a critical gap: models lack the strategic procedural knowledge to employ visual aids effectively—specifically, the decision-making of *when to draw, what to draw, and how to leverage the visualization* for subsequent deduction. Crucially, such constructions significantly reduce problem-solving difficulty, particularly when diagrams are inherently complex or benefit from intrinsic properties (Chervonyi et al., 2025).

To address this, we conducted an ablation study by injecting oracle auxiliary aids in different modalities. Our findings confirm that single-modality auxiliary aids—whether consisting solely of textual instructions or visual diagrams—cannot serve as a lossless substitute for the interleaved provision of construction statements and corresponding images, as they fail to fully encapsulate the synergistic information inherent in the multimodal context. Furthermore, we observed that this visual feedback correlates strongly with a reduction in reasoning perplexity (PPL), mirroring a *"cognitive epiphany"* where correctly constructed auxiliary lines drastically reduce the uncertainty of the subsequent reasoning trajectory.

Building on these insights, we propose a framework for Visual-Text Interleaved Chain-of-Thought. We first introduce **GeoAux-Bench**, a benchmark comprising 4,334 geometry problems and 8,470 diagrams. As illustrated in Figure 1, it establishes a precise interleaved mapping where each textual construction (T_{aux}) is explicitly paired with its corresponding ground-truth visual update (I_{aux}). To effectively leverage this data, we present **Action Applicability Policy Optimization** (A^{2}PO), a reinforcement learning paradigm designed to master strategic visual construction. A^{2}PO employs a Tri-Partition Sampling strategy to construct counterfactual reasoning paths (mandatory vs. prohibited). Based on these baselines, we employ Adaptive Reward Shaping to orchestrate the reasoning process: (1) A **Timing Reward** to discern the necessity of auxiliary lines; complemented by (2) a **Quality Reward**, grounded in reasoning perplexity, to ensure constructions genuinely simplify the solution path. During inference, we implement Visual Re-prompting to dynamically inject auxiliary diagrams, enabling the model to reason in a truly interleaved manner.

In summary, our contributions are as follows:

1. We present **GeoAux-Bench**, the first geometry benchmark to explicitly associate textual auxiliary construction steps with corresponding auxiliary diagrams.
2. We provide empirical evidence that interleaved visual-textual auxiliary representations significantly outperform single-modality counterparts by up to **1.97%**. We further demonstrate that high-quality auxiliary constructions act as an **entropy reducer**, narrowing the solution search space and lowering reasoning uncertainty.
3. We propose **A^{2}PO**, a reinforcement learning paradigm utilizing **Adaptive Reward Shaping**. By strictly regulating the *timing* and evaluating the *quality* of visual aids, A^{2}PO achieves a maximum performance gain of **3.51%** over GRPO and unconditional reinforcement strategies.

2 Related Work

2.1 Benchmarks for Multimodal Mathematical Reasoning

Visual-mathematical reasoning benchmarks, from foundational datasets (Lu et al., 2021, 2022) to recent suites like MMMU (Yue et al., 2023) and MathVista (Lu et al., 2023), have advanced MLLMs (Wang et al., 2024; Zhang et al., 2024; Kazemi et al., 2023). However, they rely on static pairs, lacking step-by-step visual demonstrations for dynamic reasoning. While Shi et al. (2025) explored interleaved formats, such data remains scarce. GeoAux-Bench bridges this gap via explicit text-visual alignment to provide dense supervision.

2.2 Visual Chain-of-Thought (VCoT)

Textual Chain-of-Thought (Wei et al., 2022; Fang et al., 2025b,a) dominates symbolic reasoning but fails to capture spatial dynamics for geometry. Visual CoT addresses this by integrating visual synthesis into deduction, with two primary streams: (1) **Tool-augmented approaches** (Hu et al., 2024; Wang et al., 2025b; Zheng et al., 2025) suffer from procedural rigidity, treating diagramming as a static step rather than a fluid cognitive strategy; (2) **Intrinsic generation models** (Shi et al., 2025; Li et al., 2025a) are prone to structural hallucinations and lack geometric precision. Our framework synthesizes these paradigms via Visual Re-prompting,

combining symbolic precision with dynamic reasoning feedback.

2.3 Reinforcement Learning for Reasoning

Reinforcement learning is central to reasoning breakthroughs (DeepSeek-AI et al., 2025). Algorithms like GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and VAPO (Yue et al., 2025) excel in text, vision (Meng et al., 2025), and tool-use (Li et al., 2025c; Hong et al., 2025b), but applying RL to geometry is non-trivial—intermediate visual constructions lack verifiable outcomes. Existing methods like GeometryZero (Wang et al., 2025c) optimize timing via tool priors (Li et al., 2025c) but overlook reasoning efficacy. Our A^{2}PO addresses this via Adaptive Reward Shaping, using contrastive sampling for timing and reasoning perplexity for utility assessment, ensuring constructions actively facilitate solutions.

3 The GeoAux Benchmark and Reasoning Dynamics

In this section, we introduce **GeoAux**, a benchmark tailored for **Interleaved-Modal Chain-of-Thought (VCoT)**. To capture the dynamic nature of geometric reasoning, GeoAux explicitly aligns textual auxiliary instructions with their corresponding visual updates ($T_{aux} \leftrightarrow I_{aux}$). We first detail the benchmark construction pipeline, followed by a pilot study that quantitatively validates the cognitive synergy between these modalities.

3.1 GeoAux-Bench Construction

To foster research into active visual-textual reasoning, we construct **GeoAux-Bench**, consisting of two subsets: the expert-annotated **GeoAux-Core** and the adapted **GeoAux-Canvas**.

GeoAux-Core. We curated geometric problems predominantly requiring auxiliary constructions from secondary school curricula and Olympiad competitions (post-January 2024) to minimize data contamination. Crucially, we restructure the solution trajectory into an interleaved format. We introduce a dedicated token pair <aux>...</aux> to encapsulate the Auxiliary Construction Instruction (T_{aux}). The closing tag </aux> explicitly marks the insertion point for the corresponding Auxiliary Diagram (I_{aux}). This structure acts as a transition operator mapping the initial visual state (I_{orig}) to an updated state (I_{aux}):

$$\mathcal{M} : (I_{orig}, T_{aux}) \rightarrow I_{aux} \quad (1)$$

Here, T_{aux} contains explicit directives (e.g., "*Connect* AB"), while I_{aux} renders these elements visually. A representative sample is shown in Figure 1. The subset is stratified into four difficulty levels: Curriculum-Junior/Senior and Olympiad-Junior/Senior.

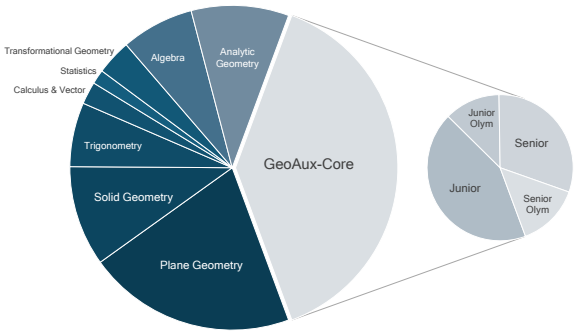
GeoAux-Canvas. To assess generalization at scale, we adapted samples from *MathCanvas-Bench* (Shi et al., 2025). We filtered for construction-heavy problems and applied our annotation pipeline to generate corresponding (T_{aux}, I_{aux}) pairs. This subset retains fine-grained subject tags (e.g., Analytic Geometry, Trigonometry), enabling detailed capability analysis.

Quality Control and Standardization. We enforced a rigorous three-stage pipeline: (1) Solvability Verification: Using Gemini-2.5-Pro (Comanici et al., 2025) to ensure problem conditions are sufficient for a unique solution; (2) Symbolic Normalization: Parsing all mathematical expressions into a unified LaTeX format; (3) Visual Enhancement: We utilized Seedream 4.0 (Chen et al., 2025b) for super-resolution to enhance image quality, followed by normalization to 512×512. Furthermore, to enable deterministic evaluation of open-ended proofs, we reformulate proof problems (e.g., "Prove $A = B$") into verifiable computational problems (e.g., "Find the ratio A/B").

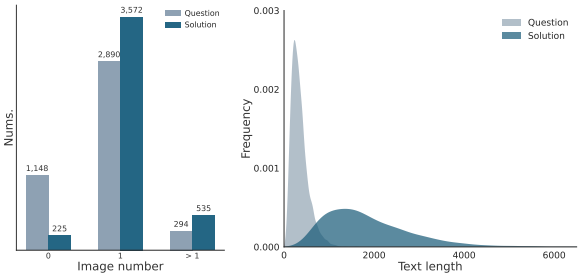
Statistics. GeoAux-Bench comprises 4,334 problems (6,523 queries) and 8,470 geometric diagrams, with an extremely high image-to-problem ratio as illustrated in Figure 2. Detailed statistics are provided in Appendix B.

3.2 Pilot Study: Complementarity of Auxiliary Modalities in Interleaved Geometric Reasoning

We conceptualize geometric construction as a critical Interleaved Reasoning Step with dual representations: the **Auxiliary Construction Instruction** (T_{aux}) and the **Auxiliary Diagram** (I_{aux}). While recent unified MLLMs (e.g., Chameleon-7B (Team et al., 2024), Bagle-7B-MoT (Deng et al., 2025)) attempt end-to-end interleaved generation, their capability to precisely render geometric constraints remains nascent. Consequently, prior works often resort to compromises: they convert diagrams that ought to be presented via the visual modality into



(a) Subject & Difficulty Distribution



(b) Image Count (c) Text Length Distribution

Figure 2: **Overview of GeoAux-Bench.**

single-modal formal languages or drawing code, thereby circumventing the challenge of visual generation.

This raises a fundamental research question: *Does T_{aux}—the textual modality describing the visual diagram—fully encapsulate the latent spatial information contained in the corresponding visual diagram I_{aux}, or do these two modalities offer complementary cognitive grounding?* To disentangle the contributions of the semantic instruction and the corresponding visual diagram, we design a controlled ablation study. Let $\mathcal{Q}$ denote the problem text and I_{orig} the initial diagram. We evaluate performance under four experimental settings:

- **Standard Settings:** $(\mathcal{Q}, I_{orig})$. Baseline reasoning without hints.
- **Textual-Only Settings:** $(\mathcal{Q}, I_{orig}, T_{aux})$. Providing only the auxiliary construction instruction.
- **Visual-Only Settings:** $(\mathcal{Q}, I_{orig}, I_{aux})$. Providing only the auxiliary diagram.
- **Interleaved Settings:** $(\mathcal{Q}, I_{orig}, T_{aux}, I_{aux})$. Simulating an ideal interleaved reasoning step with both modalities.

Perceptual Saliency Control (Visual Enhancement). A potential confounder in the *Visual-Only Setting* is visual saliency: standard (typically dashed) auxiliary lines may be too subtle for detection, leading to false negatives from *perceptual misses*.To boost the saliency of these elements and prevent them from being overlooked in complex geometric configurations, we introduce a Visual Enhancement Protocol, as illustrated in Figure 3. Specifically, we identified 200 hard samples where models succeeded with the *Textual-Only Settings* but failed with the *Visual-Only Settings*; to these, we applied the enhanced annotation (denoted $I_{\text{aux}}^{\text{red}}$) to ensure visibility, yielding two control settings: **Enhanced Visual-Only** $(\mathcal{Q}, I_{orig}, I_{\text{aux}}^{\text{red}})$ and **Enhanced Dual-Modal** $(\mathcal{Q}, I_{orig}, T_{aux}, I_{\text{aux}}^{\text{red}})$.

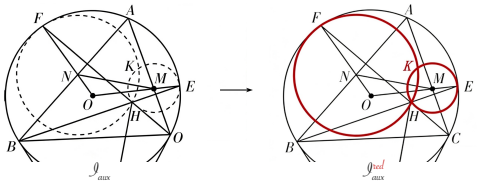


Figure 3: **Visual Enhancement Protocol.** Bold red lines are used to highlight auxiliary elements ($I_{\text{aux}}^{\text{red}}$).

Analysis of Reasoning Dynamics. Table 1 presents the performance metrics. By analyzing the intra-group trends, we derive three critical insights into the cognitive mechanism of auxiliary construction:

Setting	Input Modality	Acc% (Δ)	Tokens	PPL
G1: Text-Only Settings				
Standard	$\mathcal{Q}, I_{orig}$	23.68	1336.2	1.1250
Text-Only	$+T_{aux}$	24.24 (+0.56)	1241.3	**1.1185**
G2: Visual-Only Settings				
Visual-Only	$+I_{aux}$	23.34 (-0.34)	1258.0	1.1340
Enhanced Visual	$+I_{\text{aux}}^{\text{red}}$	24.44 (+0.76)	1303.6	1.1310
G3: Interleaved Settings				
Dual-Modal	$+T_{aux} + I_{aux}$	24.90 (+1.22)	**1230.7**	1.1350
Enhanced Dual	$+T_{aux} + I_{\text{aux}}^{\text{red}}$	**25.31** (+1.63)	1270.5	1.1323

Table 1: Reasoning Performance under Different Modal Input Combinations. Experiments are grouped by context structure for fair intra-group comparisons.

(1) **Irreplaceability and Complementarity of Modalities.** While the *Text-Only Setting* outperforms the baseline, it still falls 1.07% short of the *Interleaved Setting*. This directly answers our research question: Textual instructions (T_{aux}) cannot fully replace visual diagrams (I_{aux}). The two modalities provide complementary cognitive grounding: text defines operational instructions to guide focus on target elements and resolve semantic ambiguity, while diagrams deliver spatial relationships to clarify geometric connections and reduce uncertainty. Optimal reasoning performance

arises only when the model "sees" the spatial realization of its intent.

(2) **Lower PPL Correlates with Higher Accuracy Across Modal Settings.** Across all groups, we observe a strong positive correlation between reduced Perplexity and improved reasoning accuracy under the same modal configuration—notably, with no significant correlation between accuracy and generated token length. This aligns with human geometric reasoning intuition: just as a well-constructed auxiliary line triggers an "epiphany" that simplifies a complex problem, valid auxiliary information lowers the model's predictive entropy. A lower PPL may signals a clearer, less ambiguous solution path, which empirically boosts the likelihood of correct reasoning—an insight we leverage in our reward shaping design.

(3) **Visual Saliency Matters.** Comparing *Visual-Only* with *Enhanced Visual* (and *Interleaved* analogously), enhancing visual saliency consistently reduces PPL and improves Accuracy (up to +1.10%). Notably, this performance gain stems from modifying only 200 specific samples. This sensitivity indicates the model's performance is highly reliant on the perceptual clarity of auxiliary elements. We hypothesize visual feature enhancement enables the model's visual system to more reliably detect and anchor geometric configuration changes. This confirms visual perceptibility is a strict prerequisite for geometric reasoning. For qualitative analysis of attentional patterns, see Appendix A.

4 Method

4.1 Preliminaries: GRPO

We adopt Group Relative Policy Optimization (GRPO) as our optimization backbone. GRPO eliminates the separate value function by estimating the baseline from the group average.

Formally, given a question q, the policy generates a group of outputs $\{o_i\}_{i=1}^{G}$. The advantage A_i is estimated by normalizing group rewards:

$$A_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G}) + \epsilon}, \tag{2}$$

where ϵ is a small constant. GRPO optimizes the surrogate objective, averaged over tokens:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q,\mathbf{o}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\Big(r_{i,t}A_i,\ \text{clip}(r_{i,t}, 1-\varepsilon, 1+\varepsilon)A_i\Big) - \beta\mathbb{D}_{KL}(\pi_\theta||\pi_{\text{ref}})\right)\right], \tag{3}$$

where $r_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}$ is the probability ratio, and $\mathbf{o} = \{o_i\}_{i=1}^{G}$ denotes the sampled group.

4.2 Overview of A^{2}PO

Building upon GRPO, we propose **Action Applicability Policy Optimization (A^{2}PO)**. While standard GRPO optimizes solely for outcome correctness, geometric reasoning requires mastering the strategic applicability of auxiliary constructions—specifically, discerning whether a visual modification is beneficial for the specific problem configuration. To explicitly model this applicability, A^{2}PO introduces a **Tri-Partition Sampling** mechanism. As illustrated in Figure 4, instead of sampling from a single policy distribution, we partition the rollout trajectories into three distinct subsets. This structure enables the construction of counterfactual baselines—comparing scenarios where auxiliary lines are enforced versus prohibited—to derive an **Adaptive Reward** that guides the model on *when* to construct (Timing) and *how* to construct efficiently (Quality).

4.3 Tri-Partition Sampling with Visual Re-prompting

To quantify the marginal utility of auxiliary constructions, we sample trajectories via three distinct conditioning protocols, aggregating them into a rollout set $\mathcal{G} = \{O^+, O^-, O\}$. For a given query q and initial diagram I_{orig}, we sample N trajectories for each subset. Prompts for each protocol are provided in Appendix E:

- **Mandatory Subset (O^+):** We employ Prefix Forcing to enforce geometric intervention. Given that the model is fine-tuned during the supervised warm-up to encapsulate auxiliary commands in <aux> tags, we pre-populate the generation prefix with <aux>. This strictly forces a valid auxiliary construction step while retaining the standard prompt context.
- **Prohibited Subset (O^-):** We impose a Hard Constraint to disable visual updates. In addition to appending a negative constraint to the system prompt, we explicitly mask the logits of the <aux> and </aux> tokens during decoding, guaranteeing a reasoning trajectory devoid of auxiliary constructions and forcing the model to rely solely on the initial visual context.
- **Natural Subset (O):** The model samples autonomously using the same standard prompt as

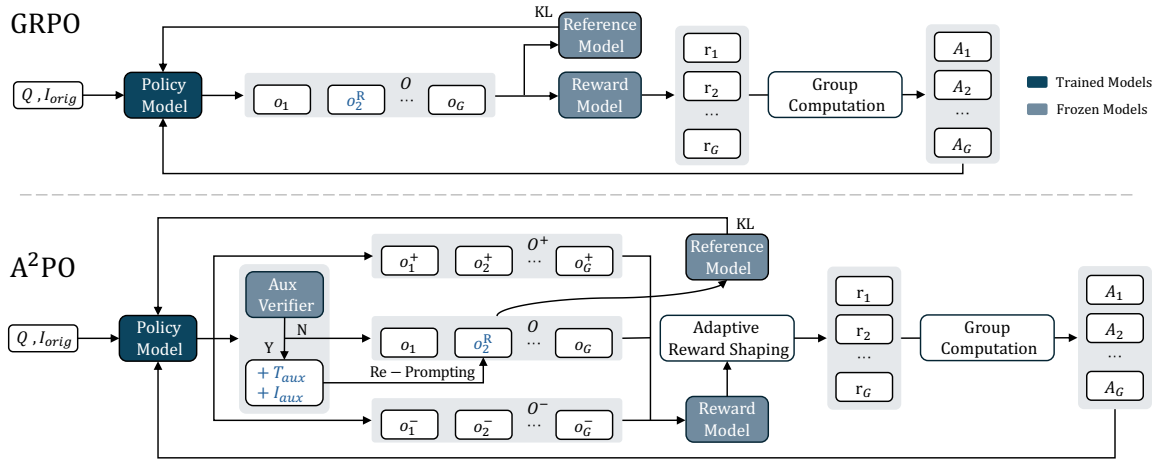


Figure 4: The framework of Action Applicability Policy Optimization (A^{2}PO). The upper panel shows standard GRPO, while the lower panel illustrates our tri-partition sampling and adaptive reward shaping mechanism.

G^+, without any intervention. This subset represents the target policy to be optimized.

Visual Re-prompting. To simulate interleaved visual reasoning despite current rendering limitations, we employ a retrieval-based injection strategy within the Natural Subset. Upon detecting a completed auxiliary command, an Aux Verifier evaluates its semantic equivalence to the ground truth. If and only if the construction is equivalent to the ground truth, we trigger re-prompting: the generation is paused, and the model is queried again with an augmented context that appends a structured "Hint" containing the ground-truth instruction (T_{aux}) and the corresponding auxiliary diagram (I_{aux}). This provides high-fidelity visual feedback contingent on correct reasoning. Prompt transformations are detailed in Appendix E.

4.4 Adaptive Reward Shaping

We design a composite reward function $R(o)$ specifically for the natural subset O:

$$R(o) = w_1 r_{acc} + w_2 r_{fmt} + w_3 r_{time} + w_4 r_{qual} \quad (4)$$

where $w_{(\cdot)}$ are weighting coefficients. While r_{acc} and r_{fmt} align with standard GRPO protocols, we introduce r_{time} and r_{qual} to optimize auxiliary construction efficacy:

4.4.1 Timing Reward (r_{time})

This component trains the model to discern the strategic necessity of auxiliary constructions. Let $\mathbb{I}_{aux}(o)$ be the auxiliary indicator and $\Delta = \mathbb{E}_{O^+}[r_{acc}] - \mathbb{E}_{O^-}[r_{acc}]$ be the utility gap. We formulate $r_{time}(o)$ with a significance margin τ:

$$r_{time}(o) = \mathbb{I}_{aux}(o) \cdot \begin{cases} 1 & \text{if } \Delta > \tau \\ -1 & \text{if } \Delta < -\tau \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

This restricts auxiliary construction to scenarios yielding a net performance gain.

4.4.2 Quality Reward (r_{qual})

Building on our pilot findings (Sec. 3.2), we posit that effective auxiliary constructions function as **entropy reducers**, mirroring the "cognitive epiphany" that manifests as reduced reasoning PPL.

We establish a perplexity baseline $\bar{\mathcal{P}} = \mathbb{E}_{O^+}[\text{PPL}]$ derived from the mandatory subset. The quality reward grants a bonus solely to valid, low-entropy auxiliary usage:

$$r_{qual}(o) = \mathbb{I}_{aux}(o) \cdot r_{acc}(o) \cdot \mathbb{I}\big(\text{PPL}(o) < \bar{\mathcal{P}} + \delta\big) \quad (6)$$

where δ compensates for the PPL overhead induced by visual re-prompting. This reward explicitly favors confident reasoning paths that simplify the solution space. Detailed hyperparameter configurations are provided in Appendix F.

5 Experiments

We split our experiments into two distinct parts to rigorously validate our contributions: (1) conducting a systematic evaluation of popular MLLMs on GeoAux-Bench to establish a difficulty baseline; (2) focusing on our proposed A^{2}PO framework, with comparisons against strong RL baselines across multiple benchmarks to verify the efficacy of our adaptive reward shaping.

5.1 Benchmark Results

We conduct a comprehensive evaluation on GeoAux-Bench involving various popular MLLMs, including proprietary SOTA models (Comanici

Model	Think	Curriculum		Olympiad		Total
		Senior	Junior	Senior	Junior	
Closed-source MLMMs						
Gemini-2.5-Pro	✓	**83.91**	**84.75**	**82.28**	88.24	**83.16**
Gemini-2.5-Flash	✓	80.10	81.68	64.56	82.13	79.53
GPT-5	✓	71.48	84.02	79.32	**88.41**	80.62
GPT-4o	✗	19.57	24.29	57.38	42.51	28.13
Claude-Opus-4.1	✓	60.36	52.70	58.23	58.45	55.82
Claude-Sonnet-4.5	✓	59.20	52.00	55.70	60.39	55.03
Doubao-seed-1.6	✓	62.02	61.42	72.57	86.47	65.00
Open-source MLLMs						
Qwen3-VL-235B-Ins.	✗	61.36	**77.80**	78.90	91.79	**74.85**
Qwen3-VL-235B-Thk.	✓	**64.68**	50.36	**89.45**	**95.17**	62.25
Qwen3-VL-30B-Ins.	✗	61.36	70.38	75.53	89.37	70.25
Qwen3-VL-30B-Thk.	✓	55.72	49.64	83.97	93.24	58.75
Qwen2.5-VL-Ins.	✗	23.05	28.81	65.82	52.17	33.25
InternVL3.5-8B	✗	38.31	26.63	54.85	66.67	36.26
InternVL3-8B	✗	25.87	30.02	67.93	55.56	35.17
GLM-4.1V-Thk.	✓	31.67	23.57	56.12	53.14	31.76
GLM-4.5V	✗	45.27	24.62	72.15	82.61	40.24
MiMo-VL-7B-SFT	✗	50.41	42.13	71.73	80.68	50.87
MiMo-VL-7B-RL	✓	50.75	41.65	71.73	80.68	50.70
United MLLMs						
BAGEL-7B-MoT	✗	7.30	7.34	34.60	38.16	12.95
BAGEL-Zebra-CoT	✗	6.80	6.94	31.65	35.27	12.03
MathCanvas-7B	✗	18.41	18.89	46.84	51.69	24.63

Table 2: Comparison on **GeoAux-Bench-Core**. Best **closed** and **open** scores are highlighted. "Ins." and "Thk." denote Instruct and Thinking models.

et al., 2025; OpenAI, 2025; Hurst et al., 2024; Anthropic, 2025a,b; ByteDance Seed Team, 2025), open-weights baselines (Bai et al., 2025a,b; Zhu et al., 2025; Wang et al., 2025a; Hong et al., 2025a), and native unified models capable of interleaved image-text generation (Shi et al., 2025; Deng et al., 2025; Li et al., 2025a). All models are evaluated under a unified setting (see Appendix F for hyperparameters). The performance on GeoAux-Bench-Core is reported in Table 2. Results on the GeoAux-Bench-Canvas subset are provided in Appendix B.

Analysis. The results highlight three critical observations:

(1) **Performance Gap & Difficulty.** A substantial performance chasm exists between top-tier proprietary models (e.g., Gemini-2.5-Pro at 83.16%) and typical open-weights baselines (e.g., InternVL3.5-8B at 36.26%). Furthermore, the benchmark proves exceptionally difficult for Native Unified MLLMs (e.g., BAGEL < 13%), which perform the worst among all paradigms. Despite being intuitively aligned with human "thinking-while-drawing," qualitative analysis (see Appendix D.1) reveals that these models fail to execute precise geometric edits, leading to severe visual hallucinations that derail subsequent reasoning.

(2) **The Analytic Shortcut.** Models frequently outperform on Senior geometry compared to Junior geometry. Case studies in Appendix D.2 reveal that MLLMs exhibit a strong preference for *Analytic Geometry* (e.g., establishing coordinate systems) to bypass visual intuition. While this algebraic conversion works for high school problems, it is often inefficient or inapplicable to the pure geometric logic required in Junior tasks, explaining the performance inversion.

(3) **Signs of Memorization.** Two anomalies point to potential data contamination rather than robust reasoning. First, models like Qwen3-VL-Thk achieve remarkable scores on the challenging Senior Olympiad split (89.45%) yet struggle significantly on the simpler Junior Curriculum set (50.36%). Second, reasoning-enhanced "Thinking" models often underperform their instruction-tuned counterparts. We attribute this to the *finite and public nature* of Olympiad problems. Unlike the vast space of curriculum exercises, high-profile competition questions are limited in number and widely circulated, making them highly susceptible to inclusion in pre-training corpora. This suggests that such performance spikes likely stem from memorizing specific problem instances rather than generalized geometric reasoning.

5.2 Performance of A^{2}PO

We evaluate our proposed A^{2}PO against SFT and strong RL baselines (GRPO (Shao et al., 2024), ToRL (Li et al., 2025c), GeometryZero (Wang et al., 2025c)). Experiments use Qwen2.5-VL across three datasets: GeoAux-Bench, external benchmarks Geomverse (Kazemi et al., 2023) and Geometry3k (Lu et al., 2021). Results in Table 3; training/inference configurations in Appendix F.

Method	GeoAux Acc%	Geomverse Acc%	Geometry3k Acc%	Avg. Acc%	GeoAux PPL ↓
Qwen2.5-VL-3B-Instruct					
SFT	23.09	56.20	39.40	39.56	**1.1389**
GRPO	**31.22**	**59.10**	50.72	**47.01**	1.1550
ToRL	28.68	58.40	47.06	44.71	1.1558
GeometryZero	29.33	57.00	**52.72**	46.35	1.1535
A^{2}PO (Ours)	**33.20**	58.40	**53.05**	**48.22**	**1.1534**
Qwen2.5-VL-7B-Instruct					
SFT	37.30	63.60	46.17	49.02	**1.0857**
GRPO	39.28	67.40	**55.49**	54.06	1.0887
ToRL	39.77	65.50	53.50	52.92	1.0941
GeometryZero	**40.18**	**68.30**	**53.72**	**54.07**	1.0945
A^{2}PO (Ours)	**42.97**	**70.70**	53.61	**55.76**	**1.0869**

Table 3: Main results on GeoAux and external benchmarks. **Best** and **Second Best** results are highlighted.

Analysis. The comparative results demonstrate

the superiority of our policy optimization strategy:
(1) **Consistent Improvements.** A^{2}PO consistently outperforms all baselines across model scales. Notably, on the 7B scale, A^{2}PO achieves an average accuracy of 55.76%, surpassing the strong GeometryZero baseline (54.07%) and standard GRPO (54.06%). This gain is most pronounced on GeoAux (+2.79% over GeometryZero), confirming that our reward shaping is particularly effective for geometric problems that heavily rely on auxiliary construction for their solution.
(2) **Reasoning Certainty & Efficiency.** A critical observation lies in Perplexity: strong baselines like GeometryZero and ToRL improve accuracy but see elevated PPL compared to SFT (e.g., up to +0.0169),suggesting performance gains may come at the cost of uncertain or convoluted reasoning. In contrast, A^{2}PO achieves the highest accuracy while maintaining a PPL nearly matching SFT. This validates Section 3.2: auxiliary constructions act as "cognitive scaffolds" reducing uncertainty. The Quality Reward guides A^{2}PO toward *elegant* simplifications—akin to a geometric "epiphany"—over convoluted computation.

5.3 Ablation Study

Table 4 presents a component-wise ablation on Qwen2.5-VL-7B-Instruct to validate the efficacy of each component.

Base Model	Method	LR	TR	QR	Vis	Acc%
Qwen2.5-7B-Instruct	GRPO	✓	✗	✗	✗	39.28
	GRPO(w/o LR)	✗	✗	✗	✗	39.52
	A^{2}PO (w/o TR, w/o Vis)	✗	✓	✗	✗	40.18
	A^{2}PO (w/o Vis)	✗	✓	✓	✗	41.17
	A^{2}PO	✗	✓	✓	✓	**42.97**

Table 4: Ablation study of A^{2}PO components, including Length Reward (LR), Timing Reward (TR), Quality Reward (QR), and Visual Re-prompting (Vis). Accuracy results are based on GeoAux-Bench.

Analysis. We analyze the incremental impact of our design choices:
(1) **Removing Length Bias (LR).** Removing the standard Length Reward yields a performance gain (39.28% $\rightarrow$ 39.52%). Unlike open-ended generation where length often correlates with detail, mathematical proofs prioritize *conciseness*. We observed that generic length incentives inadvertently encouraged verbose behaviors (e.g., repetitive reasoning). Removing this constraint allows the model to focus on the logical precision of the solution rather than artificially extending the trajectory.
(2) **Efficacy of Reward Shaping (TR & QR).** Incorporating the Timing Reward (TR) improves accuracy to 40.18%, explicitly surpassing the ToRL baseline (39.77%). While ToRL indiscriminately rewards *any valid auxiliary construction*, TR teaches the model the strategic distinction of *necessity*—rewarding the action only when it is strictly beneficial compared to a non-auxiliary path. Further adding the Quality Reward (QR) pushes performance to 41.17%. This confirms that guiding the model towards lower-perplexity (i.e., more confident) constructions effectively filters out *valid-but-redundant* auxiliary lines—constructions that are syntactically correct but offer no strategic value to the solution.
(3) **Visual Synergy (Vis).** The most significant jump occurs with **Visual Re-prompting** (+1.80%), achieving the peak accuracy of 42.97%. This empirical evidence strongly supports our core contribution: textual descriptions of auxiliary lines alone are insufficient. The model achieves optimal reasoning only when the textual construction is explicitly rendered and injected back as an updated visual context. This mechanism successfully simulates the cognitive advantage of "thinking with images" within a re-prompting framework.

6 Conclusion

In this work, we advance geometric problem solving by shifting from static perception to **active Visual-Text Interleaved reasoning**. First, we introduce **GeoAux-Bench**, the first benchmark to rigorously align textual construction instructions with corresponding visual updates, providing a precise testbed for multimodal interaction. Second, our empirical analysis reveals that visual aids function as essential **entropy reducers**: interleaved visual feedback significantly lowers reasoning uncertainty compared to single-modality inputs, validating the cognitive necessity of "thinking with images." Capitalizing on this, we propose **Action Applicability Policy Optimization (A^{2}PO)**. By employing **Adaptive Reward Shaping** to strictly regulate the *timing* and *quality* of visual interventions, A^{2}PO enables models to master the strategic deployment of auxiliary lines, achieving state-of-the-art performance. Our findings demonstrate that equipping models with the agency to actively modify their visual context is a pivotal step toward autonomous, physically grounded geometric reasoning.

Limitations

The primary limitation of this work lies in the implementation of the visual update mechanism. While our A^{2}PO framework establishes the cognitive benefits of interleaved reasoning, the current execution relies on a retrieval-based injection strategy—utilizing ground-truth auxiliary diagrams—rather than native model-generated visuals. This design choice is necessitated by the capabilities of current Unified MLLMs. Despite recent advancements, existing models still lack the fine-grained visual actuation capability required for precise geometric editing. As observed in our pilot experiments, even atomic operations (e.g., "*Connect points A and B*") frequently result in structural hallucinations or pixel-level inaccuracies that can propagate errors into the reasoning chain. Furthermore, the high inference latency associated with autoregressive image generation currently hinders efficient exploration during reinforcement learning. Consequently, our work simulates the *strategic decision-making* of visual construction rather than the *physical execution*. We posit that realizing a fully autonomous loop—where a model generates, perceives, and refines its own diagrams—requires future advancements in multimodal pre-training. Specifically, it demands a tighter alignment between high-level geometric conceptual understanding and low-level pixel editing skills. Only upon this foundation can true dynamic sampling and reinforcement learning for generative visual-text reasoning be achieved.

References

Anthropic. 2025a. Claude Opus 4.1 system card. System card, Anthropic.

Anthropic. 2025b. Claude Sonnet 4.5 system card. System card, Anthropic.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, and 46 others. 2025a. Qwen3-vl technical report.

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, and 8 others. 2025b. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923.

ByteDance Seed Team. 2025. Doubao / SEED-1.6 model series. https://seed.bytedance.com/en/seed1_6. Official description of SEED1.6 multimodal AI models.

Jiangjie Chen, Wenxiang Chen, Jiacheng Du, Jinyi Hu, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Wenlei Shi, Zhihong Wang, Mingxuan Wang, Chenrui Wei, Shufa Wei, Huajian Xin, Fan Yang, Weihao Gao, Zheng Yuan, Tianyang Zhan, and 3 others. 2025a. Seed-prover 1.5: Mastering undergraduate-level theorem proving via learning from experience.

Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, and 30 others. 2025b. Seedream 4.0: Toward next-generation multimodal image generation. *ArXiv*, abs/2509.20427.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. 2025. Thinking with generated images. *ArXiv*, abs/2505.22525.

Yuri Chervonyi, Trieu H. Trinh, Miroslav Olsák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V. Le, and Thang Luong. 2025. Gold-medalist performance in solving olympiad geometry with alphageometry2. *ArXiv*, abs/2502.03544.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3279 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaoping Nie, Ziang Song, Shi Guang, and Haoqi Fan. 2025. Emerging properties in unified multimodal pretraining. *ArXiv*, abs/2505.14683.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Xihui Liu, and Hongsheng Li.

2025a. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *ArXiv*, abs/2503.10639.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. 2025b. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *ArXiv*, abs/2509.09680.

GLM-V Team Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, and 58 others. 2025a. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. 2025b. Deepeyesv2: Toward agentic multimodal model. *Preprint*, arXiv:2511.05271.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke S. Zettlemoyer, Noah A. Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *ArXiv*, abs/2406.09403.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. Gpt-4o system card. *ArXiv*, abs/2410.21276.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *ArXiv*, abs/2312.12241.

Ang Li, Charles L. Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. 2025a. Zebra-cot: A dataset for interleaved vision language reasoning. *ArXiv*, abs/2507.16746.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vuli'c, and Furu Wei. 2025b. Imagine while reasoning in space: Multimodal visualization-of-thought. *ArXiv*, abs/2501.07542.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025c. Torl: Scaling tool-integrated rl. *ArXiv*, abs/2503.23383.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*.

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 *Annual Meeting of the Association for Computational Linguistics*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning.

OpenAI. 2025. GPT-5 system card. System card, OpenAI.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junmei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300.

Aditya Sharma, Aman Dalmia, Mehran Kazemi, Amal Zouaq, and Christopher Pal. 2024. Geocoder: Solving geometry problems by generating modular code through vision-language models. *ArXiv*, abs/2410.13510.

Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, Linjiang Huang, Si Liu, Rui Liu, and Hongsheng Li. 2025. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *ArXiv*, abs/2510.14958.

Chameleon Team, Mingda Chen, Jacob Kahn, and Shang-Wen Li. 2024. Chameleon: Mixed-modal early-fusion foundation models. *ArXiv*, abs/2405.09818.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *ArXiv*, abs/2402.14804.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 44 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *ArXiv*, abs/2508.18265.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025b. Visuothink: Empowering lvlm reasoning with multimodal tree search. *ArXiv*, abs/2504.09130.

Yikun Wang, Yibin Wang, Dianyi Wang, Zimian Peng, Qipeng Guo, Dacheng Tao, and Jiaqi Wang. 2025c. Geometryzero: Improving geometry solving for llm with group contrastive policy optimization. *ArXiv*, abs/2506.07160.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

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*, abs/2409.12122.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *ArXiv*, abs/2503.10615.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, Tiantian Fan, Zhengyin Du, Xiang Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, and 8 others. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *ArXiv*, abs/2504.05118.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *ArXiv*, abs/2505.14362.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, and 29 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv*, abs/2504.10479.

A Qualitative Analysis of Attentional Patterns

To complement the quantitative results in Section 3.2, we present the attentional patterns of Qwen2.5-VL-7B-Instruct for two representative examples before and after visual saliency enhancement (bold red auxiliary lines). No significant attentional shift toward the red-highlighted auxiliary elements is observed across cases. However, in both examples, higher attention scores are seen for the alphabetic labels of points involved in auxiliary line construction (e.g., G, M).

These observations indicate that visual enhancement may modulate attention via textual point labels rather than the auxiliary lines themselves. However, the underlying mechanism through which a minimal set of visually enhanced samples yields such substantial accuracy gains remains to be further explored.

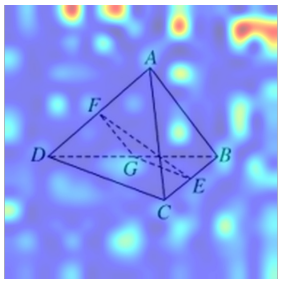
B Detailed Statistics of GeoAux-Bench

Table 5 presents the comprehensive statistical breakdown of the GeoAux-Bench dataset. The benchmark is composed of two primary subsets:

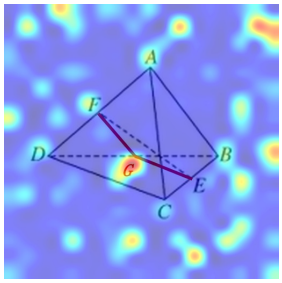
1. **GeoAux-Core:** This expert-curated subset (1,679 problems) is stratified by difficulty

Statistic	GeoAux-Core					GeoAux-Canvas	Total
	Curriculum		Olympiad		Subtotal		
	Senior	Junior	Senior	Junior			
Problems	516	722	234	207	1,679	2,655	**4,334**
Questions	647	1,288	237	207	2,379	4,144	**6,523**
Problem Diagrams	422	1,097	207	206	1,932	1,693	**3,625**
Solution Diagrams	505	892	205	199	1,801	3,044	**4,845**

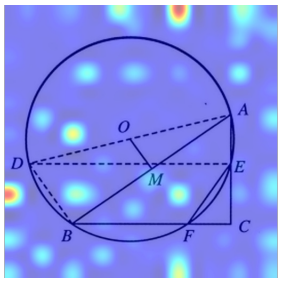
Table 5: Detailed composition of **GeoAux-Bench**. The table breaks down the dataset by subset source, difficulty level, and data modality counts.



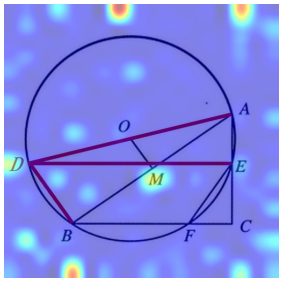
(a) Example 1: Before enhancement



(b) Example 1: After enhancement



(c) Example 2: Before enhancement



(d) Example 2: After enhancement

Figure 5: Attentional heatmaps of Qwen2.5-VL-7B-Instruct. Warmer colors indicate higher attention levels.

source, distinguishing between standard **Curriculum** problems and **Olympiad**-level competitions across both Senior and Junior grades.

2. **GeoAux-Canvas:** This subset includes 2,655 scale-augmented samples adapted from MathCanvas-Bench to enhance domain diversity.

We report metrics across four dimensions: the number of distinct *Problems*, total *Questions* (including sub-questions), original *Problem Diagrams*, and the generated *Solution Diagrams* (visual auxiliary lines).

C Results on GeoAux-Bench-Canvas

In this section, we present the supplementary evaluation results on the **GeoAux-Bench-Canvas** subset. The performance metrics for the evaluated models are directly sourced from Shi et al. (2025). Detailed breakdowns across different geometric subdomains are provided in Table 6.

D Case Study

D.1 Case of Visual Hallucinations

We present representative failure cases of Native Unified MLLMs (e.g., MathCanvas-7B) in Figure 6, as these models lack the pixel-level control required to produce accurate diagrams. This leads to a Visual-Logic Mismatch: the generated diagram contains severe distortions (e.g., curved lines, missed intersections), causing the model to hallucinate non-existent geometric properties based on the flawed visual feedback, which ultimately derails the reasoning process.

D.2 Case of the Analytic Shortcut

We present failure cases demonstrating the model's tendency to bypass geometric intuition in favor of algebraic brute force, termed as **The Analytic Shortcut**. Instead of employing geometric theorems or auxiliary lines, the model habitually establishes coordinate systems to solve problems via complex equations. As discussed in the main text, while this approach is viable for Senior high school problems, it often proves inefficient or fails to capture the pure geometric logic required for Junior geometry tasks.

E Prompt Templates

We present the prompt templates used in Tri-Partition Sampling (Section 4.3) and evaluation. Figures 8 and 9 show the templates for the Natural/Mandatory (O/O^+) and Prohibited (O^-) subsets, respectively. Figure 10 shows the template for visual re-prompting. Figures 11 and 12 display the prompts for the answer judge and auxiliary verifier.

Model	Think	Alg. (Algebra)	Ana. (Geom.)	Calc. (Vector)	Plane (Geom.)	Solid (Geom.)	Stats.	Trans. (Geom.)	Trig.	Total
Closed-source MLLMs										
Gemini-2.5-Pro	✓	68.0	**59.2**	60.2	**54.8**	**48.7**	64.5	**58.5**	**69.9**	**47.9**
Gemini-2.5-Flash	✓	63.2	56.5	54.6	40.7	40.7	61.1	46.8	64.6	39.3
Gemini-2.0-Flash	✗	39.1	32.6	38.9	31.1	25.6	51.4	28.1	38.0	21.2
GPT-4.1	✗	40.4	30.7	37.1	24.1	25.1	54.0	21.5	42.5	19.0
GPT-4.1-mini	✗	35.7	30.5	36.5	22.0	22.4	24.8	19.7	30.3	14.6
GPT-4o	✗	21.6	17.7	21.8	19.5	18.6	17.4	13.2	23.0	9.9
GPT-5	✓	**68.7**	55.5	**64.2**	45.6	36.1	64.5	42.7	66.5	43.5
Claude-Sonnet-4	✓	44.8	38.9	49.3	33.8	33.0	46.9	30.3	47.6	25.0
Seed-1.6-Thinking	✓	67.7	57.5	55.9	52.2	45.0	65.1	56.8	60.7	44.1
Qwen3-VL-Plus	✓	67.0	54.6	56.9	45.9	42.0	**66.7**	49.3	58.9	40.9
Nano-Banana	✗	55.4	50.2	51.8	34.5	36.6	56.7	39.4	60.4	33.2
Open-source MLLMs										
Qwen-2.5-VL-7B	✗	19.5	19.0	19.2	20.6	18.7	10.7	13.9	15.0	8.9
Qwen-2.5-VL-32B	✗	29.8	27.4	27.8	27.4	27.2	27.9	20.1	30.5	15.4
Qwen-2.5-VL-72B	✗	30.6	19.5	**36.4**	34.5	33.5	23.9	**33.6**	**48.9**	21.1
Gemma-3-27b-it	✗	31.3	28.4	34.4	25.8	21.0	40.0	21.0	26.9	15.8
InternVL3.5-8B	✗	32.3	**33.8**	33.8	24.2	26.9	**43.7**	16.2	14.9	16.7
InternVL3.5-30B	✗	22.2	19.9	15.1	24.9	24.3	22.1	17.4	18.4	11.7
Keye-VL-1.5-8B	✓	**33.1**	28.0	26.2	27.0	23.6	29.5	20.9	26.3	17.1
United MLLMs										
BAGEL-7B-MoT	✗	18.1	13.1	17.1	20.8	23.0	10.9	19.4	13.3	8.3
BAGEL-Zebra-CoT	✗	18.0	15.1	15.6	18.0	16.8	20.8	11.1	14.1	8.0
MathCanvas-7B	✗	29.9	27.2	17.9	40.0	35.3	23.2	29.3	40.4	21.9

Table 6: Detailed performance breakdown on the **GeoAux-Bench-Canvas** subset. Results are cited from Shi et al. (2025). Best scores in **closed** and **open** categories are highlighted.

F Implementation Details

F.1 Training Hyperparameters

Table 7 details the hyperparameters used during the Supervised Fine-Tuning (SFT) warm-up phase and the subsequent A^{2}PO training phase. We utilize `Qwen2.5-VL-7B-Instruct` as the backbone. For the SFT stage, we freeze the vision tower and the multi-modal projector, updating only the language model. For the A^{2}PO stage, we continue to freeze the vision tower to maintain visual feature stability.

Table 7: **Training Hyperparameters.** Comparison between the SFT warm-up stage and the A^{2}PO RL stage.

Hyperparameter	SFT (Warm-up)	A^{2}PO (RL)
Base Model	Qwen2.5-VL-7B-Instruct	
Precision	bfloat16	
Optimizer	AdamW	
Learning Rate	5e-5	1e-6
LR Scheduler	Cosine	Constant
Warm-up Ratio	0.1	0.0
Global Batch Size	32	24
Gradient Accumulation	4	-
Max Sequence Length	8,192	8,192
Image Resolution (min/max)	$262,144$	$262,144$
Epochs/Steps	5 epochs	650 steps
GRPO Specific Settings		
KL Coefficient (β)	-	0.01
Rollout Batch Size	-	72
Generations per Prompt (N)	-	8

F.2 A^{2}PO Algorithm Coefficients

Table 8 lists the specific weighting coefficients and thresholds defined in the reward shaping mechanism (Section 4.4).

F.3 Inference and Sampling Configuration

Table 9 details the generation parameters. We utilize a strong external model as the **Aux Verifier** and **Judge** to ensure the quality of the retrieved auxiliary diagrams and the final answer correctness.

Table 8: **A^{2}PO Reward Coefficients.**

Component	Symbol	Value
Reward Weights		
Accuracy Weight	w_{acc}	0.70
Format Weight	w_{fmt}	0.00
Timing Weight	w_{time}	0.15
Quality Weight	w_{qual}	0.15
Thresholds & Margins		
Timing Significance	τ	0.15
PPL Tolerance	δ	0.01

Table 9: **Generation & Verifier Configurations.** Settings for training rollouts, the auxiliary verifier, and final evaluation.

Parameter	Training Rollout	Evaluation
Temperature	1.0	0.0 (Greedy)
Top-p	1.0	0.0
Repetition Penalty	1.05	1.08
Max New Tokens	8,192	8,192
Accuracy Reward Model		
Verifier Model	Qwen3-30B-A3B-Thinking-2507	
Verifier Temperature	0.0 (Greedy)	
Max Tokens	24576	
Aux Verifier		
Verifier Model	Qwen3-30B-A3B-Thinking-2507	
Verifier Temperature	0.0 (Greedy)	
Max Tokens	8192	

F.4 Dataset Composition and Splits

To ensure rigorous evaluation, we strictly enforce decontamination between training and evaluation sets. Table 10 summarizes the data sources and statistics across the SFT warm-up, A^{2}PO training, and evaluation phases.

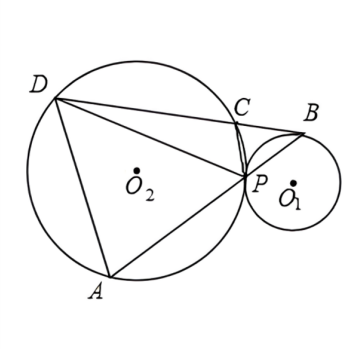
Training Data Construction. The SFT dataset is constructed using a multi-constraint instruction tuning strategy to initialize controllability. We augment training samples by creating diverse prompt-response pairs, specifically: (1) **Standard Prompts** paired with solutions containing auxiliary constructions, and (2) **Prohibited Prompts** paired with solutions strictly devoid of auxiliary lines. For the A^{2}PO stage, we employ a marginal solvability filtering strategy. We perform 10 inference rollouts per problem using the base model. We retain only samples that exhibit mixed outcomes (i.e., containing both correct and incorrect responses), explicitly discarding trivial (100% correct) or impossible (0% correct) instances to maximize gradient efficiency.

Evaluation Data. All evaluation results reported in the main paper are based on held-out test sets. For **GeoAux-Bench**, we reserve a fixed test split that is strictly excluded from all training stages.

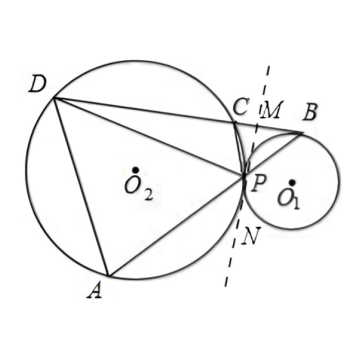
Table 10: **Data Statistics and Splits.** Detailed breakdown of sample counts. Note that for SFT and RL phases, we utilize subsets from the training splits of external datasets to avoid contamination.

Stage	Data Source / Subset	Count
Phase 1: SFT Warm-up (Mixed-Prompt)		
	GeoAux-Bench (Train Split)	1,000
	GeomVerse (Train Subset)	300
	Geometry3k (Train Subset)	300
Phase 2: A^{2}PO Training (Solvability Filtered)		
	GeoAux-Bench (Train Split)	1,500
	GeomVerse (Train Subset)	300
	Geometry3k (Train Subset)	300
Phase 3: Evaluation (Held-out Test Sets)		
	GeoAux-Bench (Test)	**1,217**
	GeomVerse (Test)	1,000
	Geometry3k (Test)	901

Question:
Two circles with centers O_1 and O_2 are externally tangent at P. From a point B on circle O_1, draw the tangent line to circle O_1, which meets circle O_2 at points C and D. The line BP meets circle O_2 again at A.\n Find the value of $\frac{BC \cdot BD}{BA \cdot BP}$.

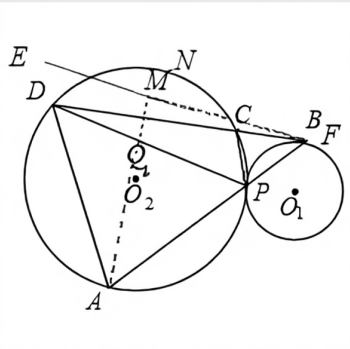


I_{orig}



I_{aux} from ground truth

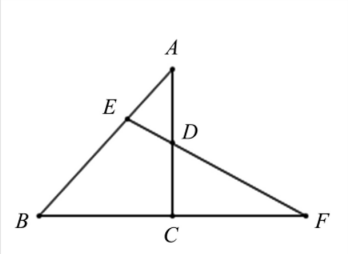
Solution:
<aux> Connect the centers O_1O_2. Extend O_1O_2 to meet circle O_2 again at E. Through E, draw the common external tangent to the two circles, touching circle O_2 at E and circle O_1 at F. This line meets the tangent from B at B at two points: above P at M and below P at N. </aux>



I_{aux} from model's output

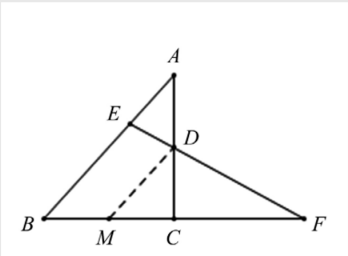
Figure 6: Failure Cases of MathCanvas-7B with Visual Hallucinations.

Question:
As shown in Figure 1, in ΔABC, $\angle ACB = \angle ACF = 90^\circ$, D is the midpoint of AC, and E is a point on AB. Connect ED. The extension of BC intersects the line ED at point F. Given $\angle F = 30^\circ$, $ED = 2$, $DF = 6$, and $EB = 2\sqrt{7}$. Then the length of BC is?



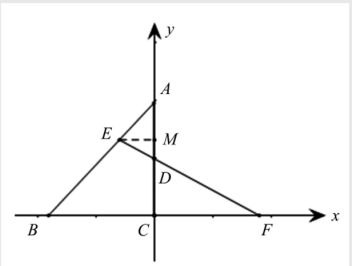
I_{orig}

Solotion From Ground Truth:
<aux> Through point D, draw $DM \parallel AB$, intersecting BC at M. </aux>
It is easy to see that $\Delta DMF \sim \Delta EBF$, yielding $\frac{MD}{BE} = \frac{DF}{EF}$, $\frac{MD}{2\sqrt{7}} = \frac{6}{8}$, which implies $MD = \frac{3\sqrt{7}}{2}$. Since $\angle F = 30^\circ$ and $DF = 6$, therefore $CD = 3$. In ΔDMC, by the Pythagorean theorem, we obtain $MC = \frac{3\sqrt{3}}{2}$. Also, since M is the midpoint of BC, we have $BC = 2MC = 3\sqrt{3}$.



I_{aux} from ground truth

Solution From Qwen2.5-VL (Analytic Shortcut):
<aux> Taking C as the origin, CF as the X-axis, and CA as the Y-axis, establish a Cartesian coordinate system. Through E, draw $EM \perp AD$,</aux>
which implies $\Delta EMD \sim \Delta FCD$, $\frac{DF}{DE} = \frac{CD}{MD}$, $\frac{6}{2} = \frac{3}{MD}$, which implies $MD=1$. By the Pythagorean theorem $EM = \sqrt{3}$, so $D(0,3)$, $F(3\sqrt{3}, 0)$. Let the equation of line FE be $y_1 = kx + b$. From $\begin{cases} b=3 \\ 3\sqrt{3}k + 3 = 0 \end{cases}$ we get $\begin{cases} b=3 \\ k = -\frac{\sqrt{3}}{3} \end{cases}$, so the equation of line FE is $y_1 = -\frac{\sqrt{3}}{3}x + 3$. When $X=-\sqrt{3}$, $Y=4$. Thus $E(-\sqrt{3}, 4)$. Let the equation of line AB be $y_2 = k_1 X + b_1$. From $E(-\sqrt{3}, 4)$, $A(0, 6)$ we get $\begin{cases} b_1 = 6 \\ -\sqrt{3}k_1 + b_1 = 4 \end{cases}$, which implies $\begin{cases} b_1 = 6 \\ k_1 = \frac{2\sqrt{3}}{3} \end{cases}$, so the equation of line AB is $y_2 = \frac{2\sqrt{3}}{3} X + 6$. Setting $y_2 = 0$, we obtain $X = -3\sqrt{3}$, $B(-3\sqrt{3}, 0)$, and $|BC| = 3\sqrt{3}$.



I_{aux} generated from text descriptions

Figure 7: An example where the model forces a coordinate system on a geometry problem, leading to unnecessary complexity.

You are a math problem solver. Please solve the following problem step by step. If the problem contains multiple subproblems, format your answers as: (1) solution to subproblem 1, (2) solution to subproblem 2...(n) solution to subproblem n.

Notice: You can construct auxiliary lines to help you solve the problem if needed. Please wrap the auxiliary line descriptions as: <aux>Connect A and B.</aux>

Question: [Here is the problem.]
[Here is the corresponding original diagrams.]

Figure 8: Standard prompt template (used for O, O^+ subsets and evaluation).

You are a math problem solver. Please solve the following problem step by step. If the problem contains multiple subproblems, format your answers as: (1) solution to subproblem 1, (2) solution to subproblem 2...(n) solution to subproblem n.

Notice: You are NOT allowed to build auxiliary lines to solve the problem. Do NOT use <aux> tags or include any auxiliary line descriptions in your response.

Question: [Here is the problem.]
[Here is the corresponding original diagrams.]

Figure 9: Prompt template for the Prohibited Subset (O^-).

You are a math problem solver. Please solve the following problem step by step. If the problem contains multiple subproblems, format your answers as: (1) solution to subproblem 1, (2) solution to subproblem 2...(n) solution to subproblem n.

Notice: You can construct auxiliary lines to help you solve the problem if needed. Please wrap the auxiliary line descriptions as: <aux>Connect A and B.</aux>

Question: [Here is the problem.]
[Here is the corresponding original diagrams.]

Hint: To solve the problem, you may consider constructing the following auxiliary line: [Here is the auxiliary construct instruction]
Here is the figure with the auxiliary lines:
[Here is the corresponding auxiliary diagrams from ground truth.]

Figure 10: Prompt template for visual-injected re-prompting.

You are an expert Mathematics Evaluator. Your task is to compare a [Model Answer] against a [Standard Answer] for specific sub-questions and judge their correctness.

Evaluation Rules:
1. **Proof Questions (Skip):**
- If the [Standard Answer] for a sub-question is exactly "{PROOF_PLACEHOLDER}", output `<ID>S</ID>` for that ID.
- Do NOT judge correctness for skipped questions.
2. **Numerical Tolerance:**
- Allow small absolute differences for decimal values (e.g., 0.25 vs 0.251 is Correct).
- Treat rounding differences as correct (e.g., 1/3 vs 0.3333).
3. **Geometric & Algebraic Equivalence (Crucial):**
- **Segments & Lines:** Treat line segments as undirected (e.g., "AB" == "BA", "Line CD" == "Line DC") unless vector notation ($\\vec{{AB}}$) is used.
- **Symmetric Relations:** Relationships like perpendicular ($\\perp$) and parallel ($\\parallel$) are symmetric.
- **Expressions:** x+y == y+x.
4. **Format & Options:**
- Multiple Choice: The letter(s) must match (A == (A)).
- LaTeX: Ignore stylistic differences (e.g., \\frac{{1}}{{2}} == 1/2).
5. **Shortcut Check:** Do NOT redo the reasoning; just compare each sub-question's final answer in the model response with the Standard Answer and judge match/mismatch.

Output Format:
- Output strictly in format: <ID>Result</ID>
- Result must be 'Y' (Correct), 'N' (Incorrect), or 'S' (Skip).
- No explanation or extra text.
- If you produce any reasoning, put it inside <think>...</think>. Outside <think>, output ONLY the <ID>Y/N/S</ID> lines.
- Do NOT solve the problem yourself; only judge correctness by comparing model outputs with the standard answers.

Task Data:
[Model Answer]:
{model_answer_text}
[Standard Answer]:
{standard_answer_text}

Figure 11: Prompt template for correctness judgment (used in evaluation and reward calculation).

You are an expert in geometry problem solving and auxiliary line construction.
Your task: judge whether the Candidate Auxiliary Line is **equivalent** to the Ground Truth Auxiliary Line.

Rules:
1. If the Ground Truth Auxiliary Line is empty, output "NO".
2. Accept minor rephrasing and different point order (e.g., "Connect A to B" == "Connect B to A").
3. The candidate must describe the same geometric construction intent (same key points/lines/relations).
4. Only judge the auxiliary construction; do NOT judge the final numeric answer.

Output Format:
Respond with only "YES" or "NO". No extra text.

Task Data:

[Ground Truth Auxiliary Line]
{gt_aux_str}
[Candidate Auxiliary Line]
{pred_aux_str}

Figure 12: Prompt template for auxiliary construction verification.