Socratic-Geo: Synthetic Data Generation and Cross-Modal Geometric Reasoning via Multi-Agent Interaction

Supplementary Material

Anonymous CVPR Submission
Paper ID 2361
CVPR 2026

A. Training Curriculum Configuration

Data Seed Construction: The initial 108 seed problems comprise geometry word problems with accompanying figures and their corresponding Python code for geometric construction. These problems were manually selected from middle school exercise collections and exam papers by human experts, with complete annotations including problem statements, figures, and solution code.

Geo170k Sampling Strategy: From the Geo170k dataset, we selected 10k problems for training. Since each geometric figure in Geo170k corresponds to 11 problems with high homogeneity, we selected one representative problem per figure to ensure data quality and diversity while also maintaining a comparable quantity to other datasets.

Stage 1: Approximately 0.4k problems, consisting of: (1) 108 manually curated seed problems, and (2) augmented problems generated by the Teacher based on Solver errors during 8-attempt solving. All augmented problems are validated for geometric correctness with ground-truth solutions.

Stage 2: Approximately 1.0k problems, consisting of: (1) all 0.4k problems from Stage 1, and (2) augmented problems generated from Solver errors during 8-attempt solving of Stage 1. All augmented problems are validated with ground-truth solutions.

Stage 3: Approximately 2.5k problems, consisting of: (1) all 1.0k problems from Stage 2, and (2) augmented problems generated from Solver errors during 8-attempt solving of Stage 2. All augmented problems are validated with ground-truth solutions.

B. Evaluation Protocol

Answer Extraction Strategy: We employ a two-tier extraction approach to identify student model responses:

- **Primary Method**: Extract answers enclosed in `\boxed{}` format using regex pattern `\boxed\{([^}]+)\}`
- **Fallback Method**: If no boxed format detected, perform full-text search for option letters (A/B/C/D) or numerical values matching the ground truth

Semantic Verification Module: An LLM-based judge (Qwen3-VL-235B) evaluates answer correctness through structured comparison:

- **Matching Rules**: Case-insensitive for multiple-choice options; unit-agnostic for numerical answers (e.g., "90" matches "90 degrees")
- **Temperature**: 0.1 for deterministic judgments
- **Max Tokens**: 10
- **Output Format**: Binary classification ("Correct" or "Incorrect")

Sampling Configuration:

- **Prompting Strategy**: Zero-shot with three randomized prompt templates to reduce variance
- **Student Model Temperature**: 0.1
- **Max Tokens**: 1024
- **Repetition**: Each problem solved N times independently (e.g., $N = 8$ for Mean@8 metric)
- **Timeout**: 300s for student model, 120s for teacher model

Mean@N Metric: A problem is considered *passed* if at least one attempt out of N repetitions is judged correct by the teacher model. The final score is computed as:

$$\text{Mean@N} = \frac{\text{Number of Passed Problems}}{\text{Total Problems}}$$

Implementation Details: Evaluations run with configurable parallelism (default: 16 concurrent workers) using ThreadPoolExecutor. Trajectory files containing all intermediate results are saved incrementally every 10 problems to ensure fault tolerance and enable result inspection.

C. Baseline Models

We compared against the following representative models from GenExam's main results:

Closed-source models: GPT-Image-1 [22], Gemini-2.5-Flash-Image [13], Imagen-4-Ultra [14], Seedream 4.0 [3], Seedream 3.0 [12], FLUX.1 Kontext max [2].

Open-source T2I models: Qwen-Image [33], HiDream-I1-Full [4], FLUX.1 dev [2], FLUX.1 Krea [2], Stable Diffusion 3.5 Large [23].

Open-source unified MLLMs: Show-o2 [34], BAGEL and BAGEL (thinking) [18], Janus-Pro [8], Emu3 [31], BLIP3o [6], BLIP3o-NEXT [7].