PRISM: Multimodal Physics and Mathematical Reasoning

Purna Chandra Sekhar V [1]

[1]IIT Delhi

Motivation

Entangled visual perception and math computation cause deterministic failures in STEM reasoning:

- **Error Propagation:** Processing both modalities in a single context allows initial perceptual mistakes to corrupt downstream math reasoning.
- **Ineffective Sampling:** Standard test-time scaling (e.g., Majority Voting) yields diminishing returns on systematic visual errors.

Objective

Our goal is to improve multimodal STEM reasoning by preventing early perceptual errors from corrupting downstream computation. We present **PRISM**, a modular agentic framework designed to:

- **Grounding:** Extracts explicit semantic entities, topologies, and spatial bounding boxes.
- **Enrichment:** Combines extracted visual structures into a unified textual problem description.
- **Program-Aided Reasoning:** Separates theoretical planning from computational execution using program-aided reasoning.

Visual System

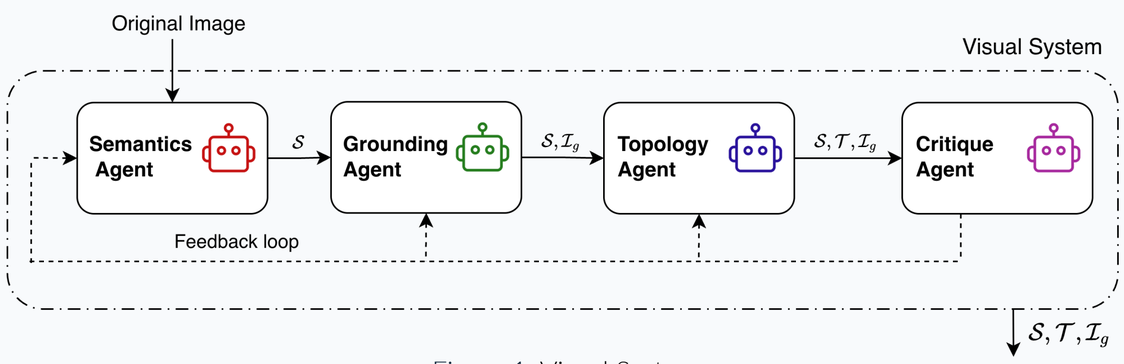

Figure 1. Visual System

This stage extracts and grounds physical entities from diagrams.

- **Extraction & Topology:** Identifies key semantic entities (e.g., masses, resistors) and explicitly maps their structural relationships.
- **Spatial Grounding:** Generates bounded images via OpenCV to precisely localize components for subsequent reasoning stages.

Enrichment System

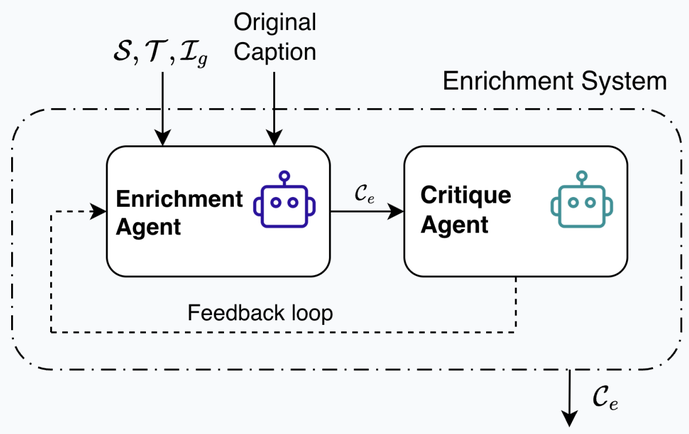

Figure 2. Enrichment System

This stage translates visual structures into a unified problem description.

- **Caption Enrichment:** Integrates grounded entities and topology into the original caption to form a complete representation.
- **Constraint Verification:** An integrated critique loop continuously catches omitted variables and hallucinated constraints.

Solver System

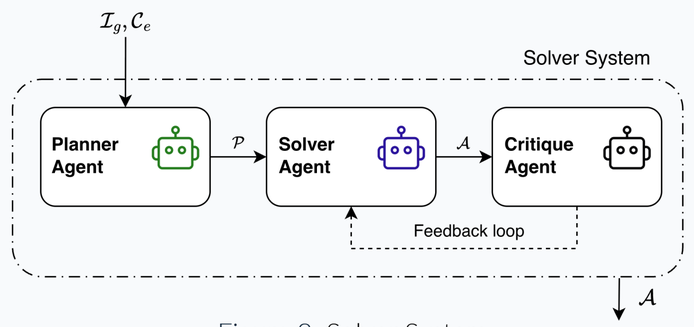

Figure 3. Solver System

This stage separates theoretical planning from deterministic execution to reliably solve problems.

- **Structured Planning:** Formulates a step-by-step execution plan, classifying each step as theoretical or computational.
- **Program-Aided Execution:** Delegates numerical calculations to SciPy, using an integrated critique loop to trace and correct errors.

Results

We evaluate PRISM on the SeePhys benchmark (the 200-question mini subset). This dataset spans eight cognitive difficulty levels, testing both complex reasoning and visual dependency. Results are reported for **Gemini 3 Flash** (Tables 1 & 2) and **Gemini 3.1 Flash Lite** (Table 3).

+17.0 pp
Gain on L8 (PhD) problems over zero-shot

81.8%
Accuracy on Vision-Essential tasks

+27.5 pp
Accuracy boost for lightweight Flash Lite

+16.7 pp
Generalization boost on the MATH-Vision mini benchmark

Table 1: Accuracy across SeePhys Difficulty (Gemini 3 Flash) (in %)

System	Overall	L4	L5	L6	L7	L8
Zero-Shot	72.0	58.7	**85.7**	**84.2**	**85.7**	61.7
Majority@5	77.5	**80.4**	**85.7**	**84.2**	71.4	61.7
PoT[1]	77.0	73.9	**85.7**	**84.2**	71.4	68.1
CAR[2]	79.0	78.3	**85.7**	**84.2**	**85.7**	68.1
PRISM	**81.0**	78.3	**85.7**	**84.2**	**85.7**	**78.7**

Table 2: Visual Dependency (in %)

System	Essential	Optional
Zero-Shot	72.3	71.2
Majority@5	75.0	**84.6**
PoT	77.0	76.9
CAR	79.0	78.8
PRISM	**81.8**	78.8

Table 3: Impact on Flash Lite (in %)

System	Overall Accuracy
Zero-Shot	37.0
Majority@5	43.5
PoT	60.5
CAR	62.5
PRISM	**64.5**

[1]PoT: Program of Thoughts • [2]CAR: Caption-Assisted Reasoning