TheoremExplainAgent: Towards Video-based Multimodal Explanations for LLM Theorem Understanding

♠ †Max Ku*, ♥Thomas Chong*, ♠Jonathan Leung, ♠ Krish Shah, ♥Alvin Yu, ♠†Wenhu Chen

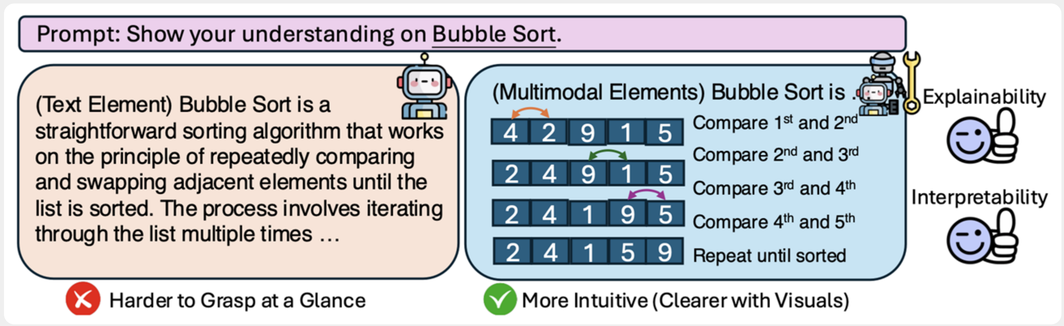
♠ University of Waterloo, ♥Votee AI, †Vector Institute

TL;DR

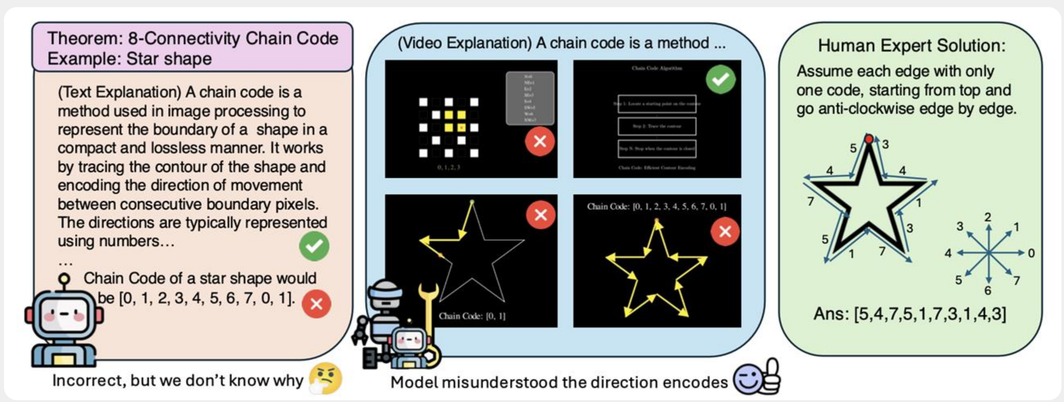
We explored an agentic approach to generate exploratory theorem explanation videos using Manim, supported by a new benchmark to evaluate **whether language models can produce visually meaningful and factually accurate multimodal explanations.** TheoremExplainAgent achieves a high success rate in generation, but the generated videos often suffer from layout issues such as misaligned elements and inconsistent visuals. Nonetheless, **we envision that future systems can generate high-quality video explanations to demonstrate theorem understanding.**

Why Multimodal Explanations Matter

A strong reasoning model should not only generate correct conclusions but also communicate them effectively. Visualization enhances human intuition by making abstract concepts more concrete and revealing hidden relationships.



We do not have knowledge of a thing until we have grasped its cause (Aristotle, 1901).

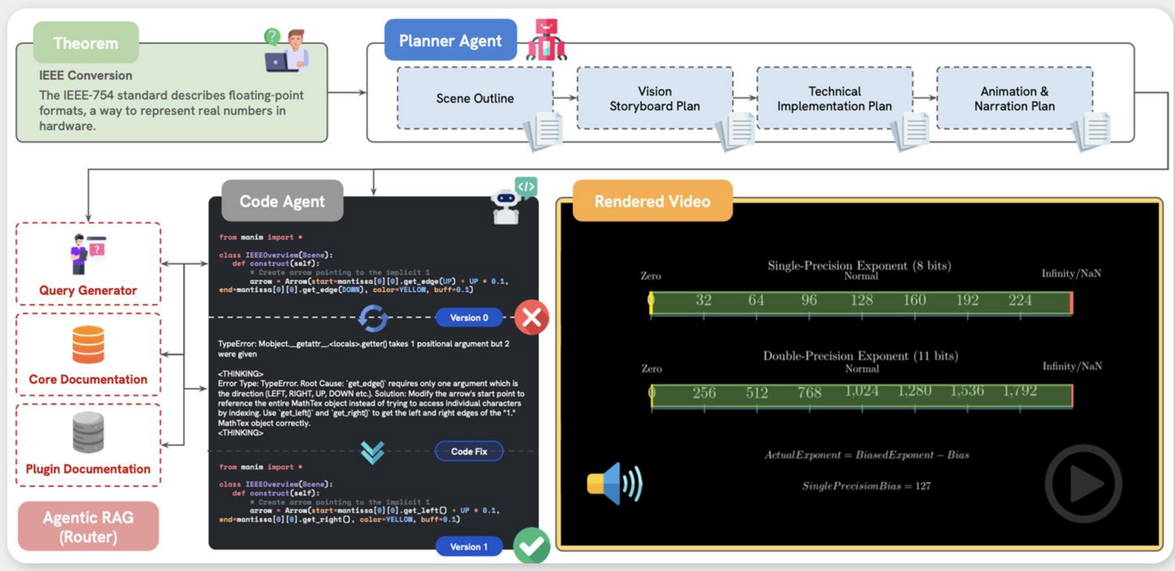


Visual explanations expose reasoning errors more clearly than text, making it easier to diagnose model mistakes.

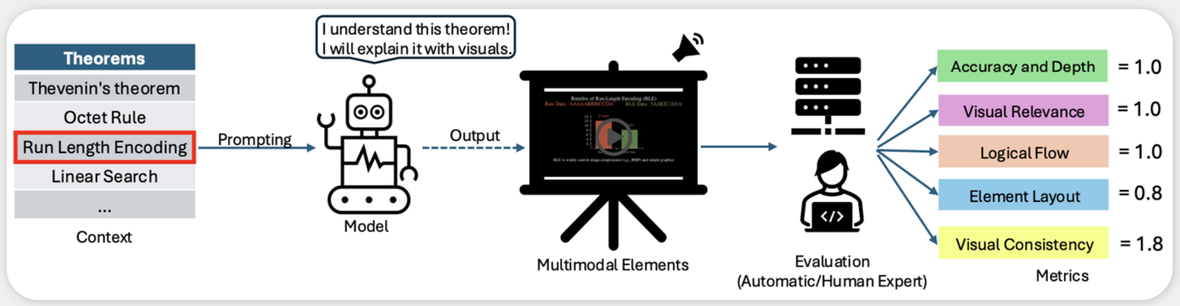
How TheoremExplainAgent works?

TheoremExplainAgent consists of two LLM agents. Taking a theorem as input, the planner agent creates plans for execution. The coding agent then generates Python scripts to produce visuals and audio. Agentic RAG system is also used to route generated queries to different documentations for better code retrieval.

Code/Data



We challenge it by introducing our TheoremExplainBench benchmark with 360 STEM theorems to see if the model can create a proper video explanation. Then we designed a novel suite of five metrics to evaluate the generated videos



So how GOOD and EXPENSIVE is TheoremExplainAgent?

Agent success rate in generating complete videos across different difficulty levels and

Agent	Easy	Medium	Hard	Math	Phys	CS	Chem	Overall
GPT-4o	61.3%	57.5%	46.2%	61.7%	55.0%	58.3%	45.0%	55.0%
GPT-4o + RAG	42.5%	57.5%	37.5%	70.0%	40.0%	41.7%	31.7%	45.8%
Claude 3.5-Sonnet v1	2.5%	1.2%	2.5%	1.7%	1.7%	1.7%	3.3%	2.1%
Claude 3.5-Sonnet v1 + RAG	18.8%	13.8%	11.2%	23.3%	10.0%	20.0%	5.0%	14.6%
Gemini 2.0-Flash	20.0%	11.2%	12.5%	16.7%	8.3%	21.7%	11.7%	14.6%
Gemini 2.0-Flash + RAG	23.8%	21.2%	16.2%	26.7%	15.0%	20.0%	20.0%	20.4%
o3-mini (medium)	**93.8%**	**91.2%**	**96.2%**	**95.0%**	**93.3%**	**93.3%**	**93.3%**	**93.8%**
o3-mini (medium) + RAG	83.8%	82.5%	80.0%	81.7%	90.0%	88.3%	68.3%	82.1%

Correlation on Metric-Human correlation and Inter-rater Agreement

	Spearman	Krippendorff's α
Accuracy and Depth	0.14	0.45
Visual Relevance	**0.72**	0.36
Logical Flow	0.16	0.56
Element Layout	**0.42**	0.31
Visual Consistency	0.17	0.36

Agent success rate in generating complete videos across different difficulty levels and subjects

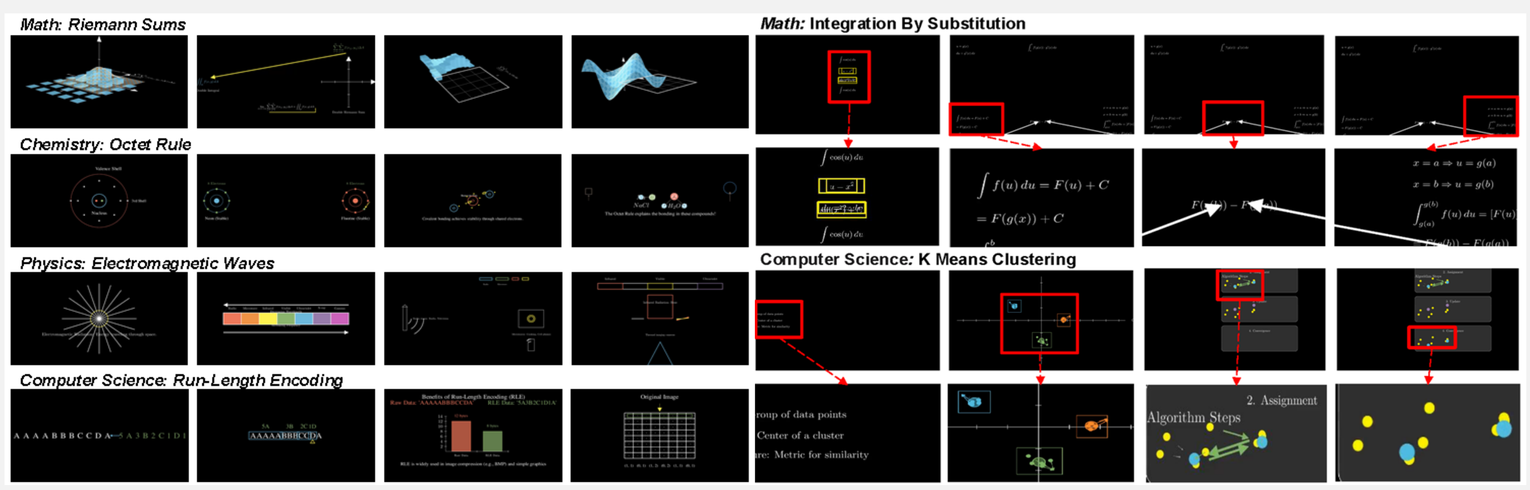
Agent	Accuracy and Depth	Visual Relevance	Logical Flow	Element Layout	Visual Consistency	Overall Score
GPT-4o	0.79	0.79	**0.89**	0.59	0.87	0.78
GPT-4o + RAG	0.75	0.77	0.88	0.57	0.86	0.76
Claude 3.5-Sonnet v1	0.75	**0.87**	0.88	0.57	**0.92**	**0.79**
Claude 3.5-Sonnet v1 + RAG	0.67	0.79	0.69	0.65	0.87	0.71
Gemini 2.0 Flash	**0.82**	0.77	0.80	0.57	0.88	0.76
Gemini 2.0 Flash + RAG	0.79	0.75	0.84	0.58	0.87	0.76
o3-mini (medium)	0.76	0.76	**0.89**	0.61	0.88	0.77
o3-mini (medium) + RAG	0.75	0.75	0.88	0.61	0.88	0.76
Human-made Manim Videos	0.80	0.81	0.70	**0.73**	0.87	0.77

Average output tokens, cost, and inference time for generating one full video.

Agent	Input Tokens	Output Tokens	Cost(USD)	Time(s)
GPT-4o	350000	84000	1.71	1120
GPT-4o + RAG	840000	84000	2.94	1260
Claude 3.5-Sonnet v1	350000	91000	2.42	2240
Claude 3.5-Sonnet v1 + RAG	1050000	101500	4.67	2380
Gemini 2.0-Flash	595000	119000	0.1	1120
Gemini 2.0-Flash + RAG	1120000	119000	0.16	1260
o3-mini (medium)	434000	154000	1.16	1680
o3-mini (medium) + RAG	945000	154000	1.72	1820

Case Studies

- Our system generates high-quality exploratory videos across various subjects, especially in easy theorems.
- Videos in these fields generally have higher visual quality than those in Chemistry
  - complex molecules and equipment are often represented by **overly simple geometric shapes**
- A common minor flaw across all videos is suboptimal layout:
  - **overlapping text, inconsistent object positioning**



Challenges

- **Fragile Code Generation:** Agent often produces buggy or non-existent Manim code, causing rendering failures.
- **Flawed Visuals:** Imperfect animations with misaligned text, overlapping shapes, and inconsistent layouts.
- **Metric Gaps:** Automated scores (narrative coherence & consistency) do not yet align well with human evaluations.