







# 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

ACL 2025 Oral

Paper ID: #733

Interpretability

# Background: What is Multimodal Explanations?

Definition: Combining text, visuals (animations), and narration to explain a concept.

- Humans often rely on visual scaffolding to explain abstract concepts
  - e.g. drawing figures on a papers, lecture videos with slides, etc.

#### Prompt: Show your understanding on Bubble Sort.

(Text Element) Bubble Sort is a straightforward sorting algorithm that works on the principle of repeatedly comparing and swapping adjacent elements until the list is sorted. The process involves iterating through the list multiple times ...

(Multimodal Elements) Bubble Sort is Explainability Compare 1st and 2nd Compare 2<sup>nd</sup> and 3<sup>rd</sup> Compare 3<sup>rd</sup> and 4<sup>th</sup> Compare 4<sup>th</sup> and 5<sup>th</sup> Repeat until sorted

Harder to Grasp at a Glance

More Intuitive (Clearer with Visuals)

# Why Multimodal Explanation Matters

## Is textual QA enough to test AI model understanding?

Models can exploit superficial cues in text-based tests (answer order, etc.), potentially overestimating true understanding

- Real-world concepts are often inherently spatial and hard to describe with solely text
- Text-only outputs can hide reasoning errors (e.g., hallucinated logic)
- Visual explanations expose flaws in logic and flow

# Multimodal Explanations for LLM Theorem Understanding

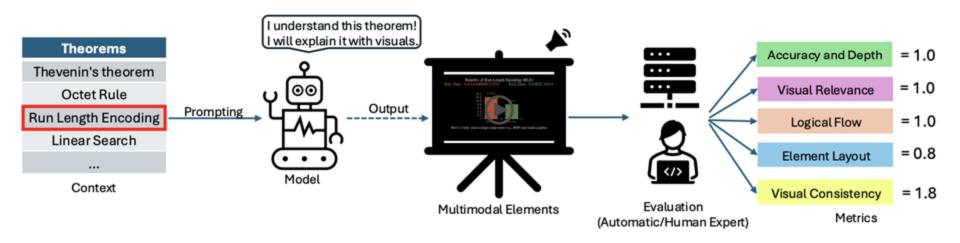
#### **Problem Definition:**

Generate >1 min video that explains a given theorem

- Long-form narrative (sound/text)
- Visual coherent structure (visual)
- Procedural accuracy (logic)
- Domain knowledge (understanding)

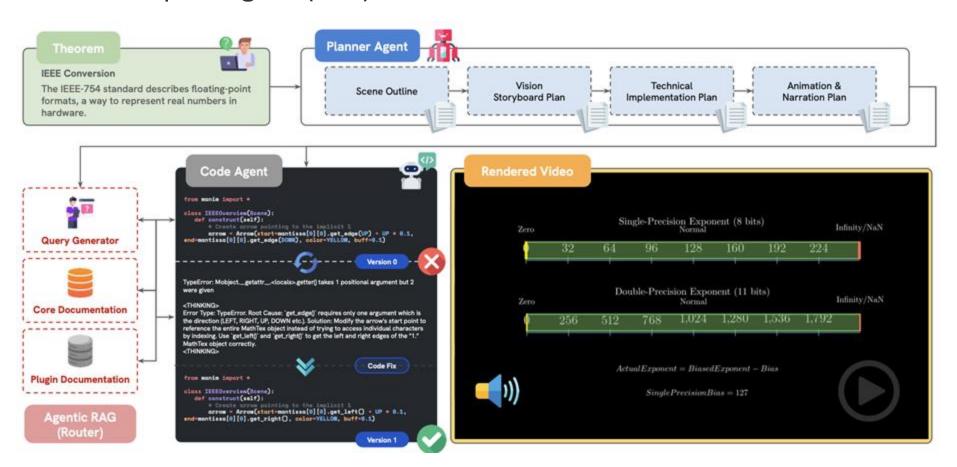
This is hard with current AI models, and we present TheoremExplainAgent as the first attempt.

# TheoremExplainAgent Framework



- Input: Theorem + Context
- Process: Agent(s) plan and generate multimodal elements (animations, narration).
- Output: Explanatory Video
- Evaluation: Assessed using automated metrics.

## TheoremExplainAgent (TEA)



# TheoremExplainBench (TEB) Dataset

A standardized benchmark to evaluate multimodal explanations.

- 240 theorems across 4 STEM disciplines
- Sourced from OpenStax, LibreTexts.
- 3 Difficulty Levels: Easy (HS), Medium (UG), Hard (Grad) -80 each.
- 5 Metrics on explanations video quality.

#### **Physics**

**Electricity and Circuits** Classical Mechanics Electromagnetism Optics Fluid Mechanics Waves and Sound Gravitation Wave Physics Astrophysics Statistical Physics Plasma Physics

Quantum Mechanics

#### Computer Science Discrete Mathematics

Boolean Algebra Algorithm Analysis **Data Structures** Computational Geometry Information Theory Operating Systems Algorithms Digital Logic Design Image Processing Signal Processing

Data Compression Machine Learning Computer Networks Computer Architecture Theory of Computation Programming Fundamentals Graph Theory Digital Signal Processing Digital Logic Dynamic Programming

Computational Complexity

Chemistry Chemical Kinetics Chemical Equilibrium Quantum Chemistry Physical Chemistry Laboratory Techniques Crystallography Organic Chemistry Inorganic Chemistry Electrochemistry Spectroscopy Acid-Base Chemistry

Thermodynamics Atomic Structure Chemical Reactions and Stoichiometry Chemical Bonding Periodic Table and Elements Separation Techniques Solid State Chemistry Analytical Chemistry Redox Chemistry Gas Laws

#### Math

Geometry Complex Analysis

Algebra Functions

Sequences and Series

Combinatorics Conic Sections

Trigonometry Calculus

Numerical Analysis

Vector Calculus Multivariable Calculus

Group Theory



## **Experimental Setup**

#### **Models Tested**

GPT-4o, Claude 3.5 Sonnet v1, Gemini 2.0 Flash, o3-mini (medium). Used for both Planner and Coder roles.

#### **Evaluation**

Among all 240 theorems in TEB:

- 1) Whether models can write proper visualization code to generate the video?
- 2) How effective are the multimodal explanations? Against actual human-made Manim videos?
- 3) Case studies to show visual explanations can expose flaws in logic and flow.

## Results: Video Generation Success Rate

Agent	Easy	Medium	Hard	Math	Phys	CS	Chem	Overall
GPT-40	61.3%	57.5%	46.2%	61.7%	55.0%	58.3%	45.0%	55.0%
GPT-40 + 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%

- o3-mini consistently outperforms others (93.8% overall success)
- **GPT-40** is moderate, struggles with complexity. **Gemini 2.0 Flash** struggles most.
- Math has highest success; Chemistry is most challenging (complex object rendering).
- RAG slightly decreased success rates in most cases (potential noise/distraction).

# Results: Video Quality Scores

(on successful videos)

Agent	Accuracy and Depth	Visual Relevance	Logical Flow	Element Layout	Visual Consistency	Overall Score
GPT-40	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

#### Insights

#### MLLM as judge:

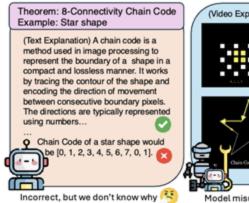
Human videos excel at Element Layout & Visual Relevance.

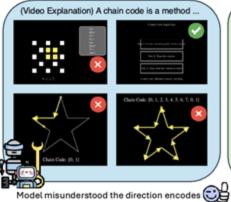
Element Layout is a common challenge for Al models (avg ~0.6).

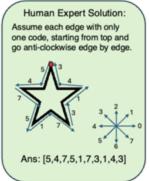
o3-mini (0.77) and GPT-4o (0.78) achieve good overall scores, comparable to human-made videos (0.77). Claude 3.5 Sonnet v1 (0.79) slightly higher but low success rate.

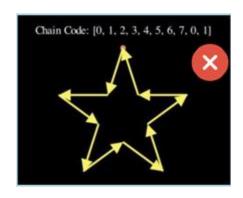
# Case Study: Visual Error Diagnosis

Multimodal explanations expose deeper reasoning flaws.









**Example: 8-Connectivity Chain Code** 

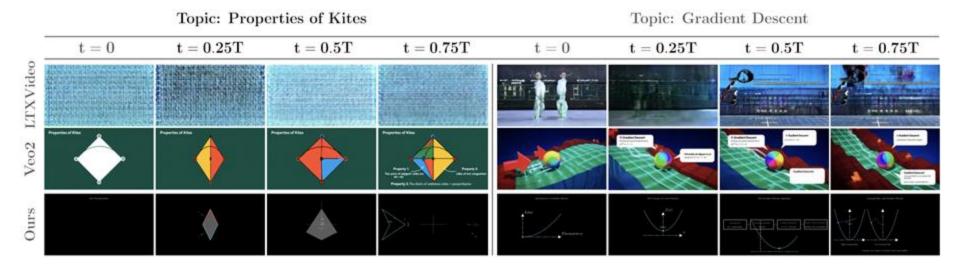
#### **Text Explanation:**

Might show the final answer is wrong, but why is unclear. Model seems to "understand" the definition but applies it incorrectly.

#### **Video Explanation:**

Clearly shows the model misunderstanding direction encoding (wrong arrows, incorrect path tracing). Makes the reasoning error explicit and diagnosable.

### Can Video Models be an alternative solution?



#### Findings:

Video models lack reasoning/planning to produce explanatory videos. However, recent video model development shows promising direction to serve as the visual (and even audio) component in the future, replacing the coding agent role.

## Conclusion

#### Conclusion

We contributed TEA (agentic system for multimodal theorem explanation videos) & TEB (benchmark/metrics).

Agentic planning is crucial for long-form coherent videos. o3-mini shows strong generation capability.

We demonstrated that Multimodal explanations are vital for deep understanding and effective error diagnosis.

However Visual element layout remains a key challenge for existing approaches, and the evaluation effectiveness is unexplored.

# Thank you!

Project Page (tiger-ai-lab.github.io/TheoremExplainAgent/), Code Available



