ICCV OCT 19-23, 2025 HONOLULU HAWAII

Bootstrapping Physics-Grounded Video Generation through VLM-Guided Iterative Self-Refinement

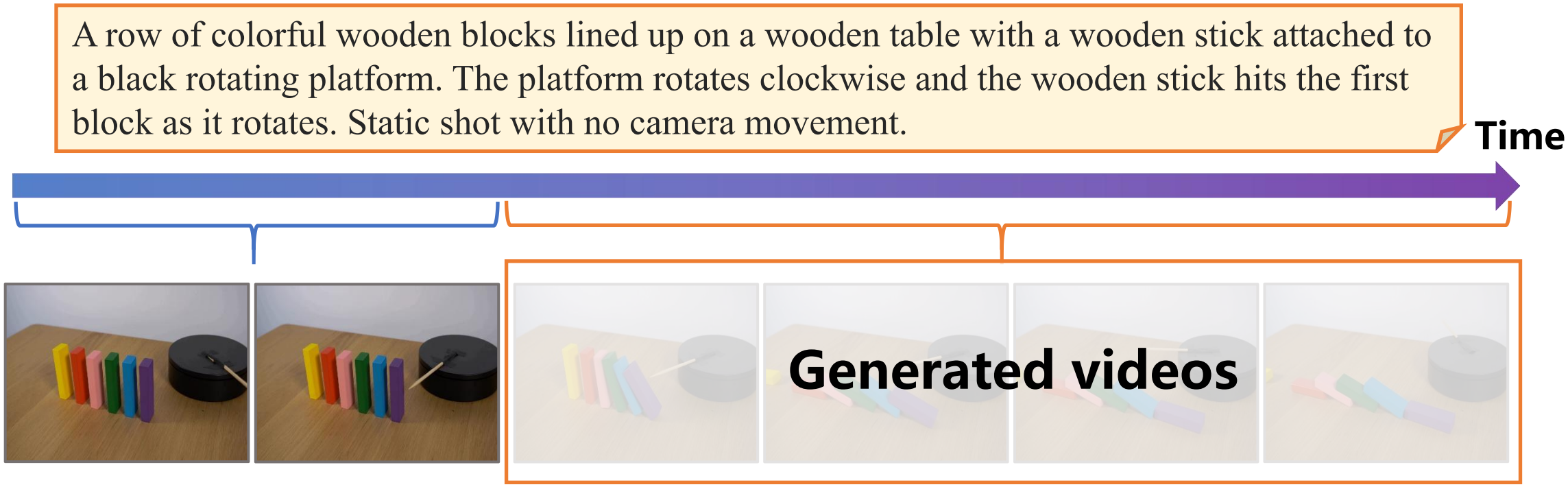
Team **MR-CAS**

Yang Liu, Xilin Zhao, Peisong Wen, Siran Dai, Qingming Huang

Oct 19, 2025

Background

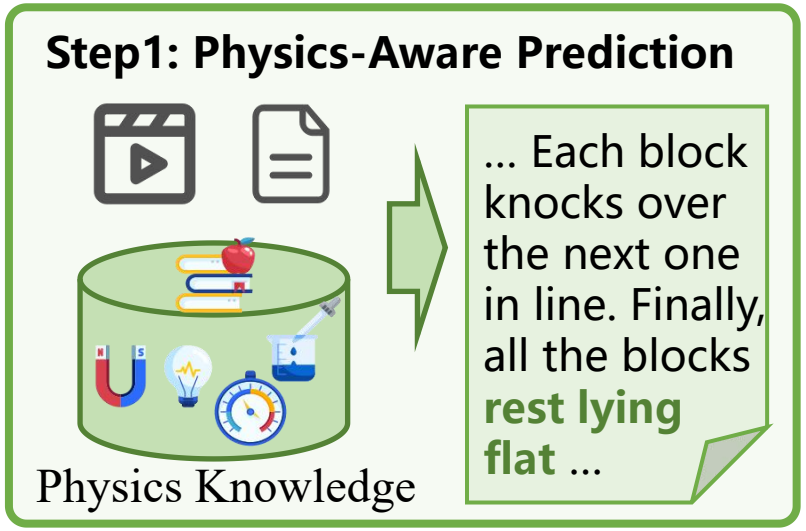
- **Overview**: Test the physical-aware ability of generation models on Physics-IQ Benchmark
- **Goal**: Generate a 5-second video that conforms to physical principles, based on a 3-second prefix video and a brief text description
- **Categories**: Solid Mechanics, Fluid Dynamics, Optics, Thermodynamics, Magnetism

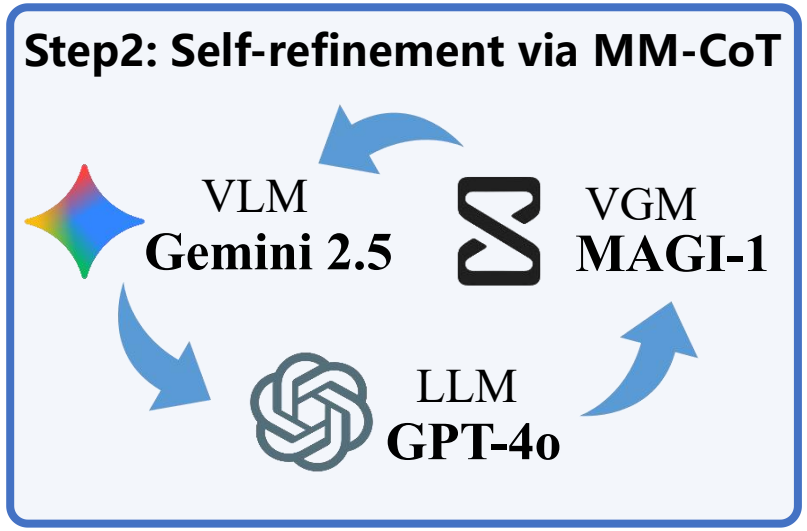


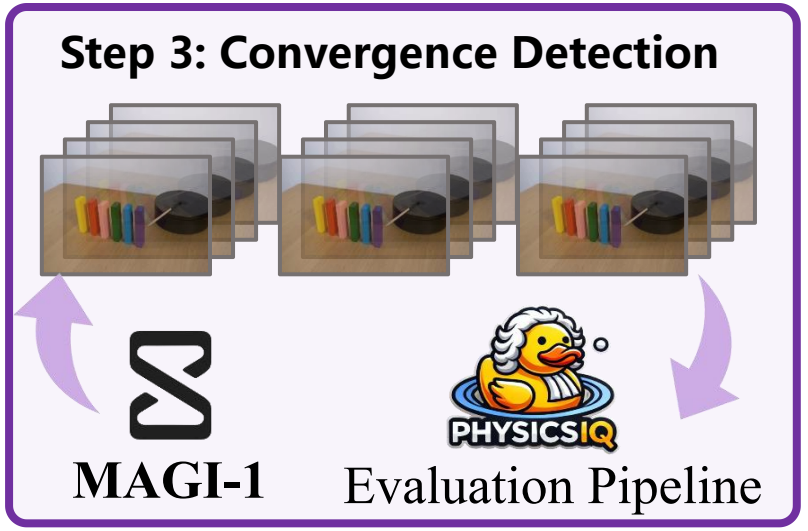
[1] Motamed S, Culp L, Swersky K, et al. Do generative video models understand physical principles?[J]. arXiv

Method

- **Motivation**
  - Q1: What are the bottlenecks in physics-aware video generation?
  - Q2: How to activate the potential of video generation models in physics conformity?
- **Methods:** VLM-guided iterative self-refinement via multimodal chain-of-thought

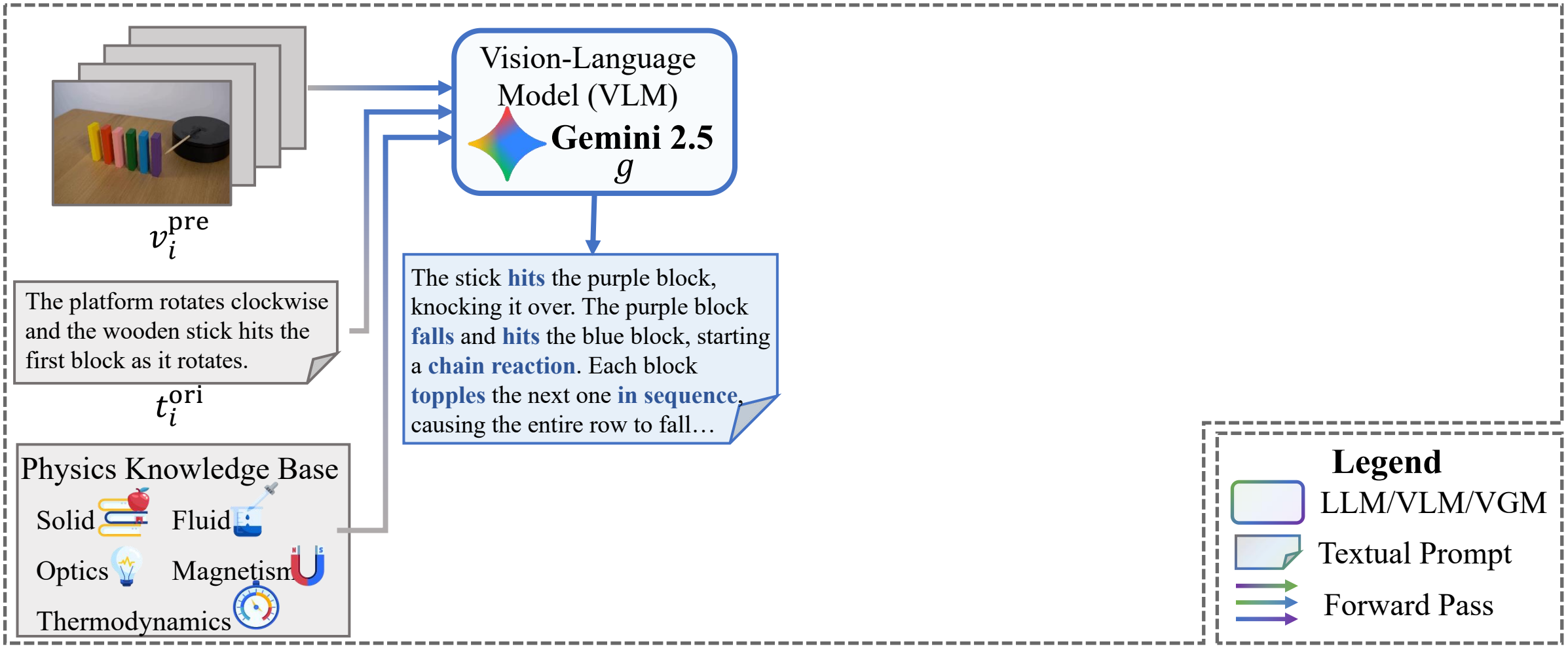






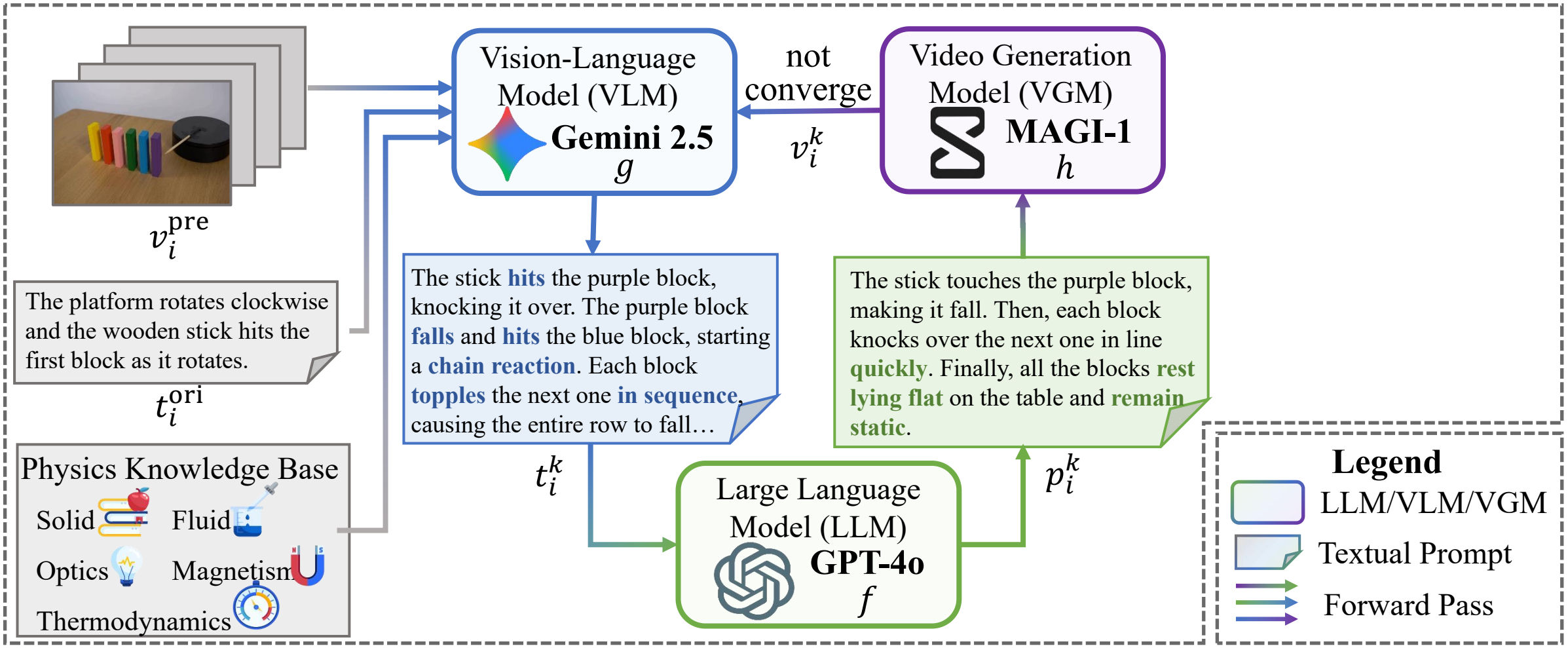
Method

➢ Step 1: Physics-Aware Text Prediction



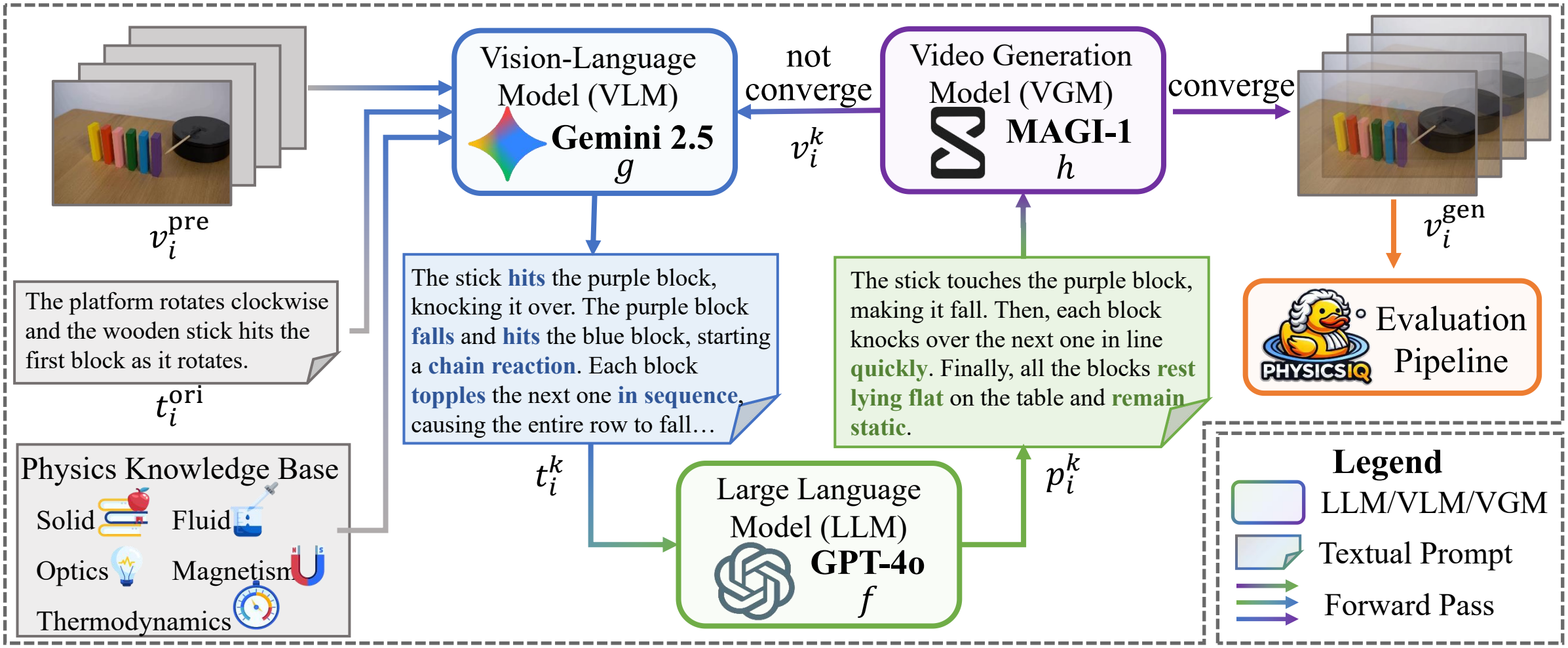
Method

➢ Step 2: Iterative Self-Refinement via MM-CoT



Method

➢ Step 3: Convergence Detection and Output



Method

➢ Example

Original	A row of colorful wooden blocks lined up on a wooden table with a wooden stick attached to a black rotating platform. The platform rotates clockwise and the wooden stick hits the first block as it rotates.	
1st Loop missing physical clues	… The rotating platform begins to turn clockwise. The attached wooden stick sweeps around, gradually approaching the row of colorful wooden blocks. **As it completes its rotation**, the stick makes contact with the first block, causing it to topple over.	
2nd Loop complex physical description	… The attached stick makes contact with the first block, transferring kinetic energy and applying a torque that overcomes the block's static **equilibrium**. **Pulled by gravity**, the first block pivots and falls, striking the next block. This impact propagates **a wave of momentum and energy** down the line, causing each successive block to topple in a chain reaction…	
3rd Loop abstract physical constraints	… The black object rotates clockwise at a steady speed, and its attached stick pushes the purple block over. The purple block falls and strikes the blue block, starting a **chain reaction**. Each block **topples in sequence**, knocking over the next one down the line from purple to yellow.	
4th Loop suitable expression	… The black object rotates clockwise at a steady speed, and its attached stick **pushes** the purple block over. Six blocks **fall one by one**. Finally, all the blocks come to rest **lying flat** on the table…	

Results

➢ Quantitative Results

- ➢ Iterative prompts leads to consistent performance gains
- ➢ The best attempt achieves a Physics-IQ score of **62.38**, which improves upon the baseline by 6.07.

Table 1. Performance across Iterative Loops on the Physics-IQ Benchmark. ∗ indicates partial refinement on incomplete prompts.

No.	Method	Infer. Steps	Physics-IQ Score (↑)
1	1st Loop	16	49.80
2	2nd Loop	16	48.31*
3	3rd Loop	16	51.65
4	4th Loop	16	52.92
5	1st Loop	32	49.49
6	4th Loop	32	49.15*
7	Ensemble {1,2,5}		57.09
8	Ensemble {1,2,3,4,5,6}		**62.38**

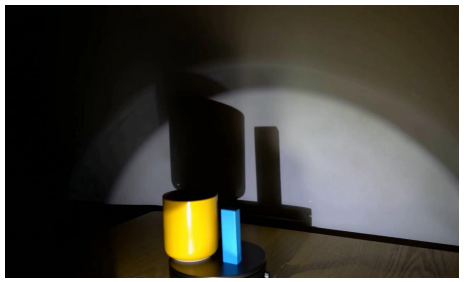
Results

➢ Qualitative Results

- ➢ For relatively **simple** physical scenarios, the video generation model produces results that closely align with real-world physical principles
- ➢ "**Simple**": slow and continuous changes, single target, and simple interactions

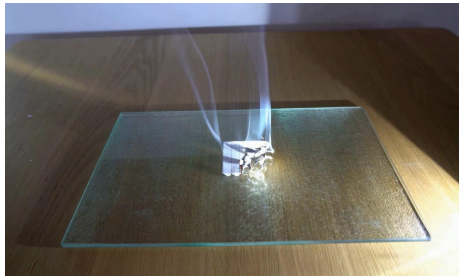
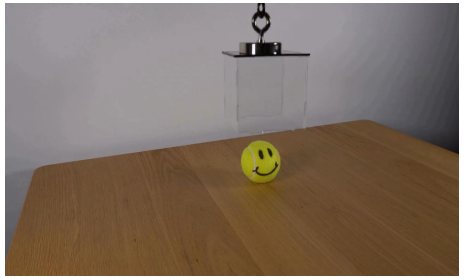
Solid Mechanics **Fluid Dynamics** **Optics** **Thermodynamics** **Magnetism**





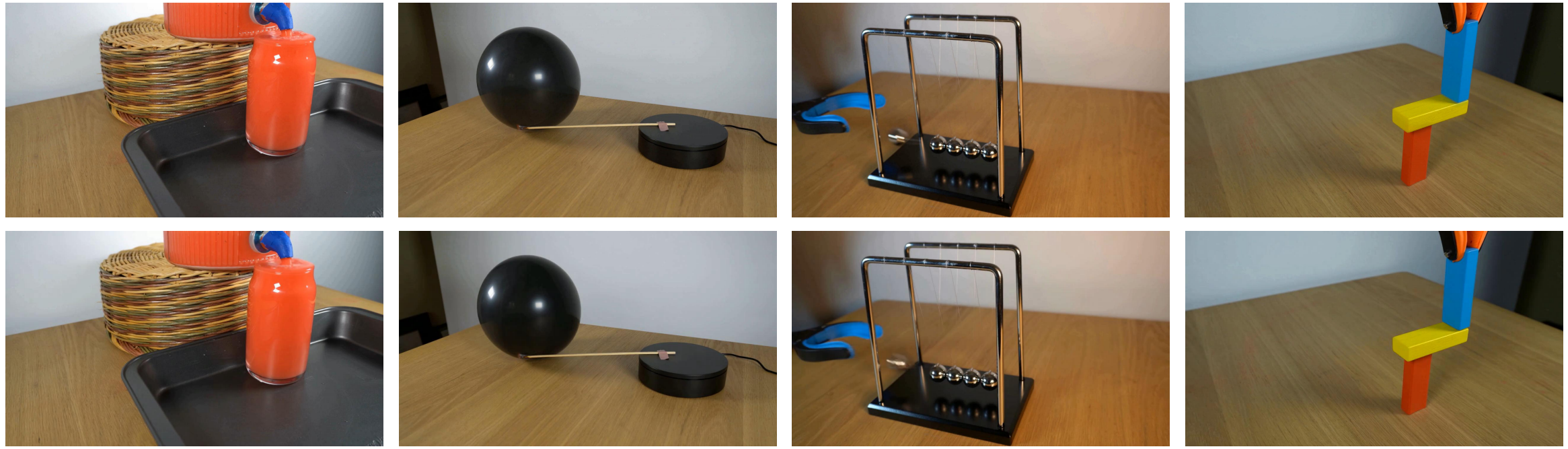




Results

- **Failure Cases Analysis**
  - Video generation models are often data-driven rather than logic-based reasoning.
  - Video generation models lack the knowledge of the real-world physical principles.



Conclusion

ICCV OCT 19-23, 2025 HONOLULU HAWAII

- **Conclusion**
  - Q1: What are the bottlenecks in video generation with physical perception?
    - The lack of large-scale data labeled with physical rules for training, this makes it hard for them to truly understand physics
  - Q2: How to activate the potential of video generation models in physics conformity?
    - Training with physics-labeled data and interaction
    - Distilling and fine-tuning representations to align with visual foundation models
    - Using advanced LLMs and VLMs to optimize prompts

Physics-aware video generation still has a long way to go…

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Thanks for your listening!

E-mail: liuyang232@mails.ucas.ac.cn

Homepage: https://yafeng19.github.io

Personal Website