Code2Video: A Code-Centric Paradigm for Tutorial Video Generation

Yanzhe Chen

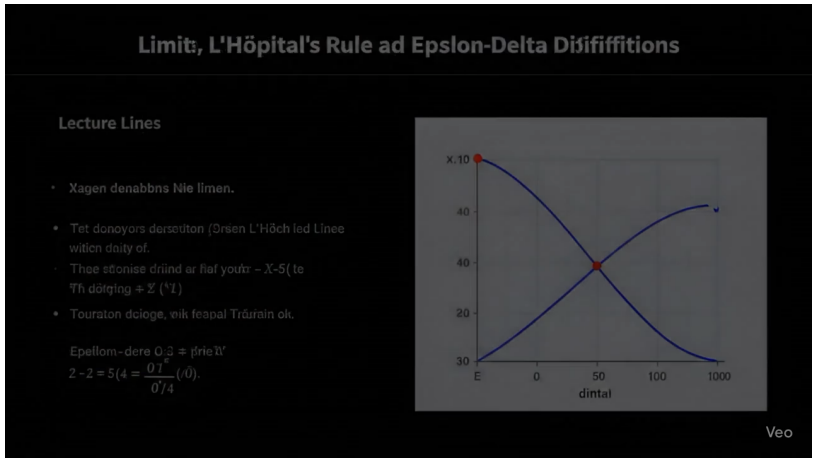
2025/09/12

Topic Introduction (1/4)

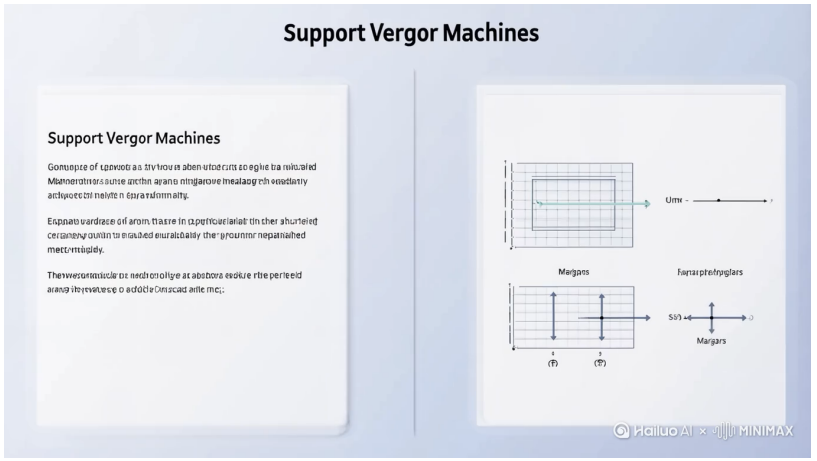
- Learning new knowledge is most effective when language and perception work together—linking explanation to experience rather than treating them in isolation
- This leads to our goal: creating tutorial videos for each concept, including lecture lines and corresponding animated explanations

Topic Introduction (2/4)

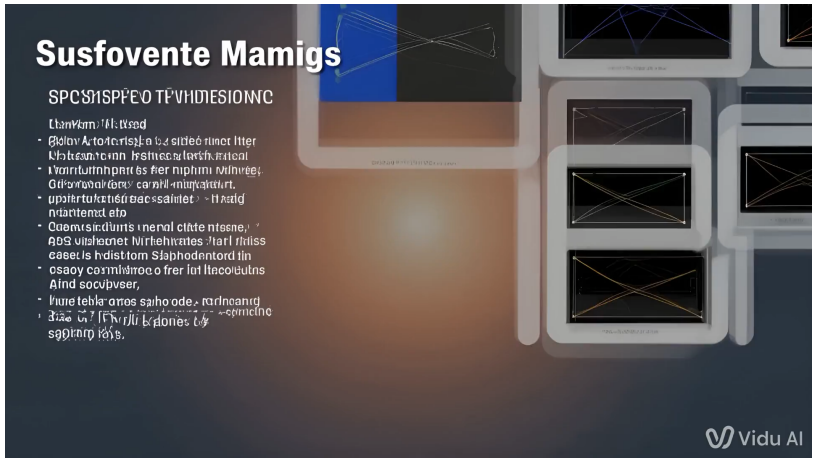
- Although recent generative models advance pixel-space video generation
- Remain limited in high-quality tutorial videos, which demand knowledge, precise visual structures, and coherent



Veo3



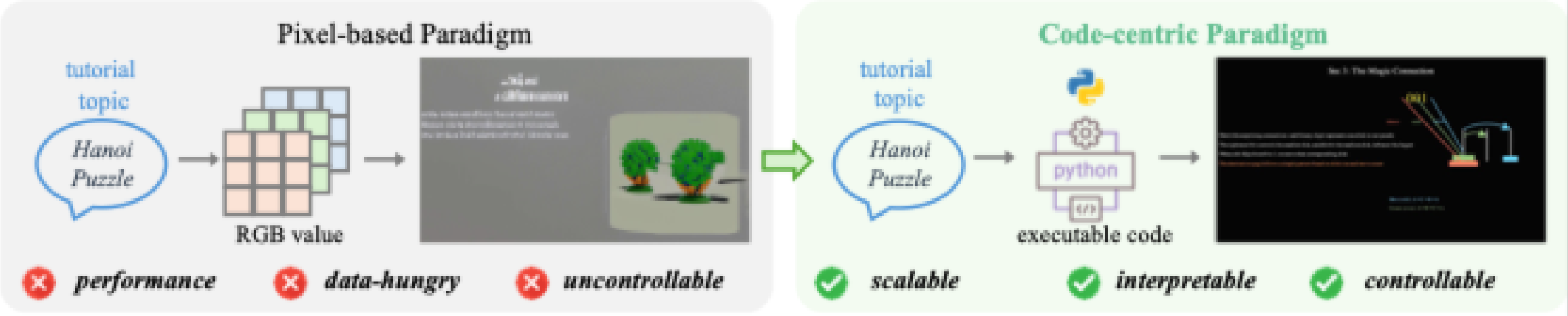
Minimax



Vidu

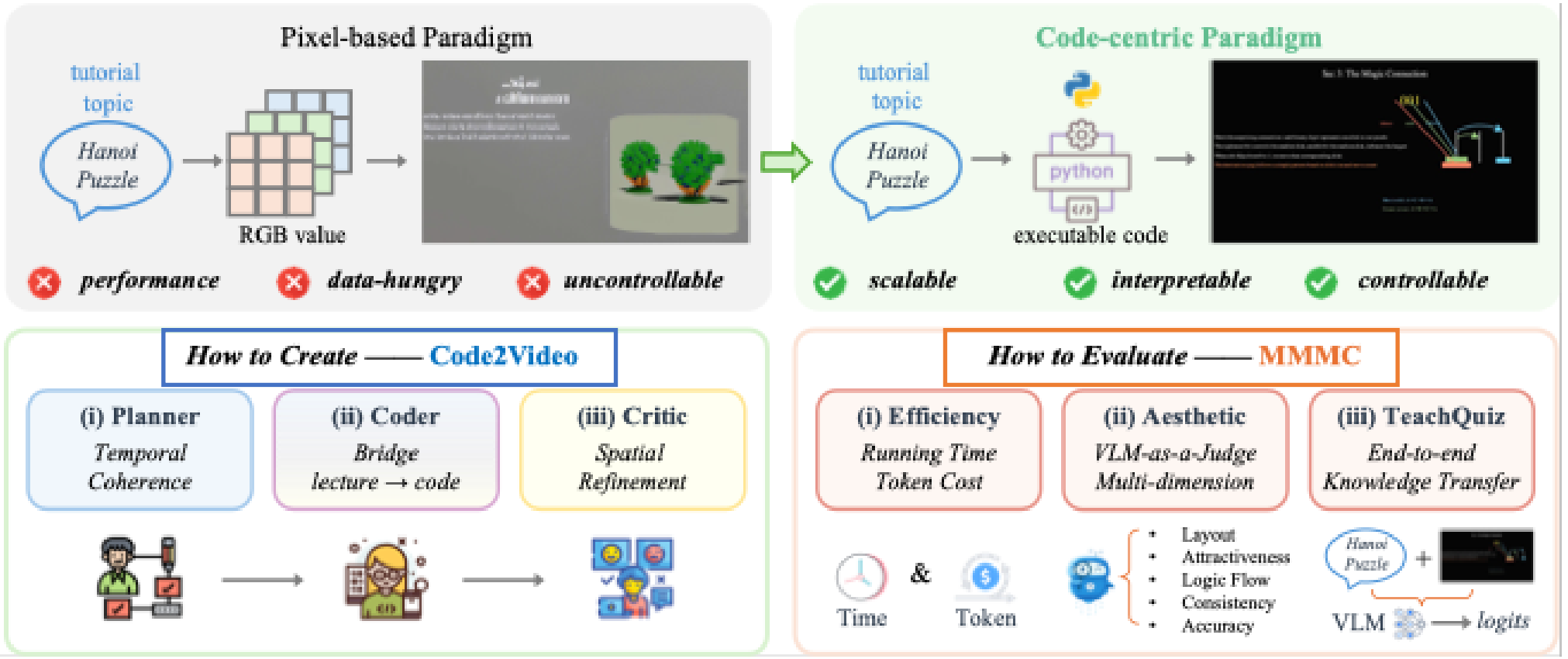
Topic Introduction (3/4)

- Such requirements are better addressed through a renderable environment, which can be explicitly controlled via logical commands (e.g., code).
- Motivation: Unlike black-box models, code-centric pipelines are scalable, interpretable, and controllable



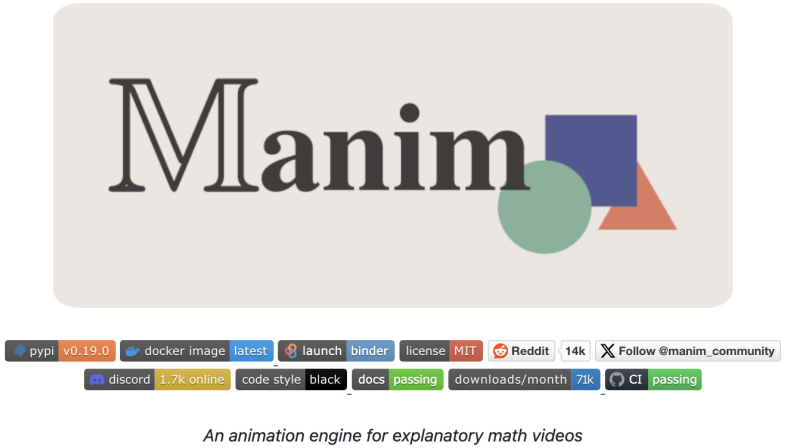
Topic Introduction (4/4)

- Two Challenges:
  - How to create high-quality videos with temporal coherence & spatial clarity ?
  - How to evaluate beyond appearance, ensuring effective & aligned ?



Benchmark (1/4)

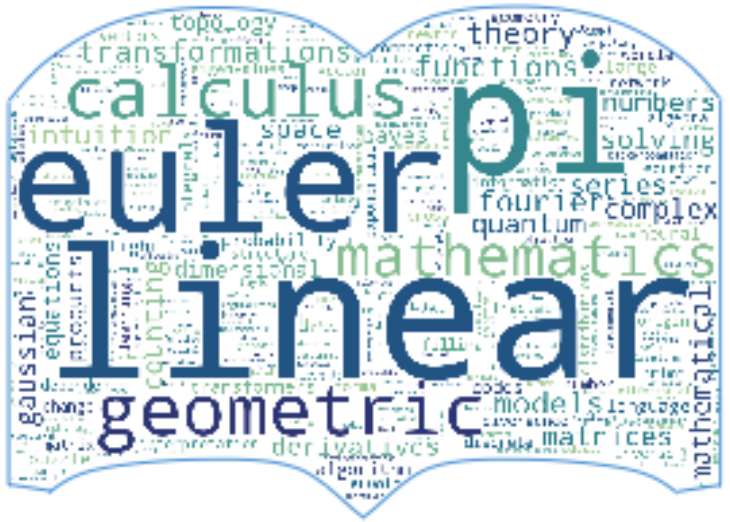
- Task Definition:
  - To transform a tutorial topic query into executable Manim code that, once rendered, produces a tutorial video
- Dataset Construction:
  - We choose Manim as the base for it offers fine-grained control, especially for math area
  - We source from 3Blue1Brown, which combines impact with expert Manim scripts



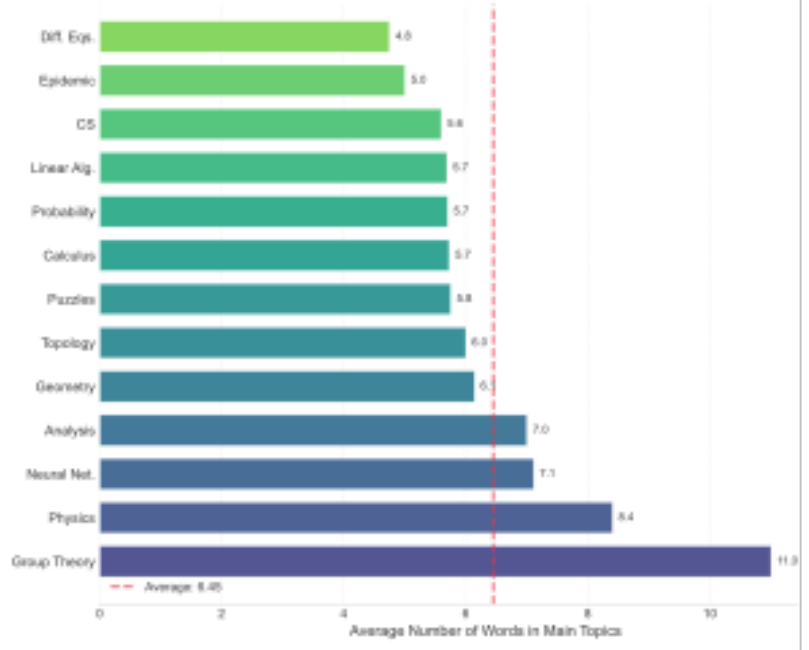
Benchmark (2/4)

- Dataset Statistics:
  - 456 tutorial videos, 117 full-length and 339 timestamped segmented clips
  - 13 high-level categories: *e.g. geometry, topology, neural networks, physics, …*

(1) Categories and Distribution



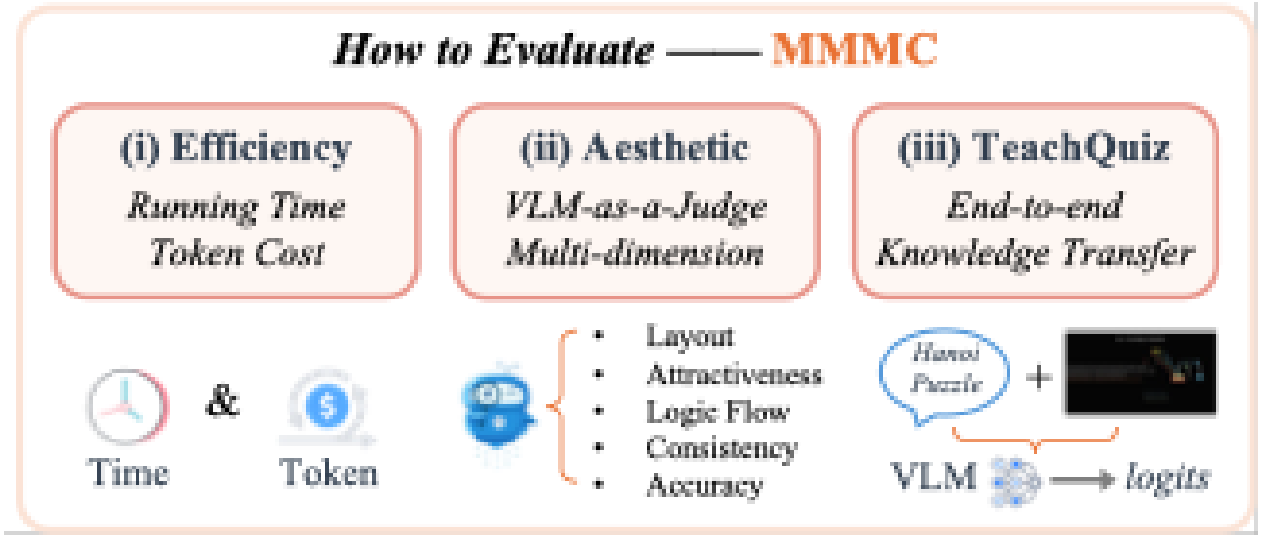
(2) Tutorial Topic Word Cloud



(3) Tutorial Topic Length

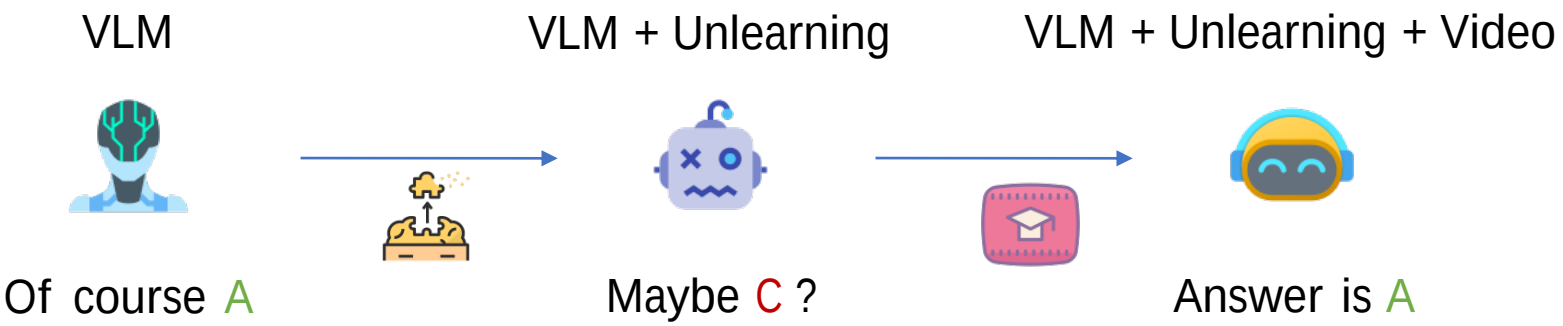
Benchmark (3/4)

- MMMC: Multi-discipline Multimodal Coding Benchmark
- *Efficiency*: along 2 axes
  - average code generation time & token consumption per video
- *Aesthetics*: VLM-as-a-Judge along 5 axes
  - *Element layout, Attractiveness, Logic Flow, Visual Consistency, Accuracy & Depth*



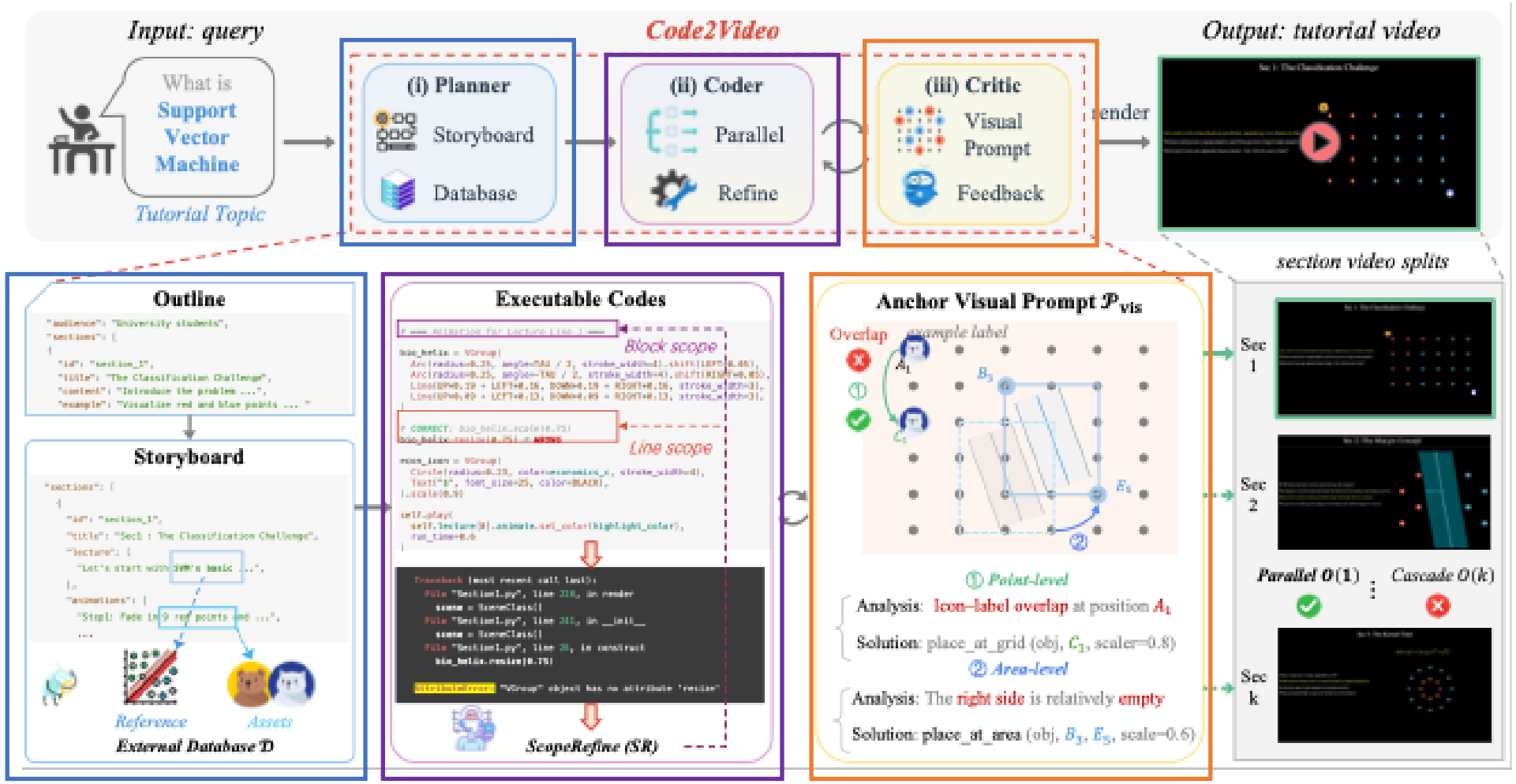
Benchmark (4/4)

Q: When a point moves on the complex plane, what *visual change indicates multiplication by i*?
A: Rotation
B: Shift
C: Disappear
D: Flash



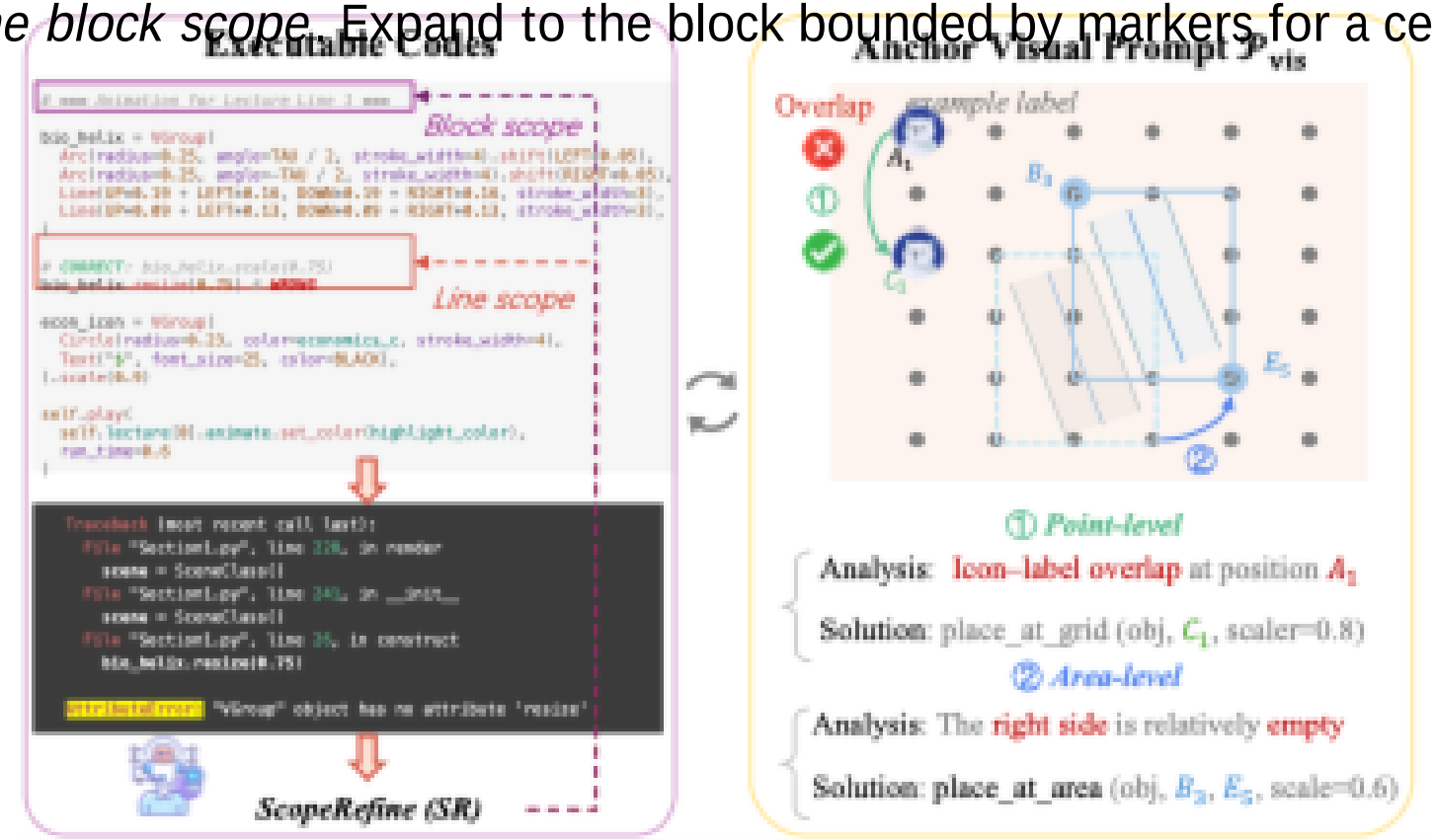
Method (1/3)

- Overall Pipeline: Planner converts a tutorial topic into a storyboard and retrieves visual assets; Coder performs parallel code generation with scope-guided refinement; Citric uses anchor visual prompts to adjust layout and clarity

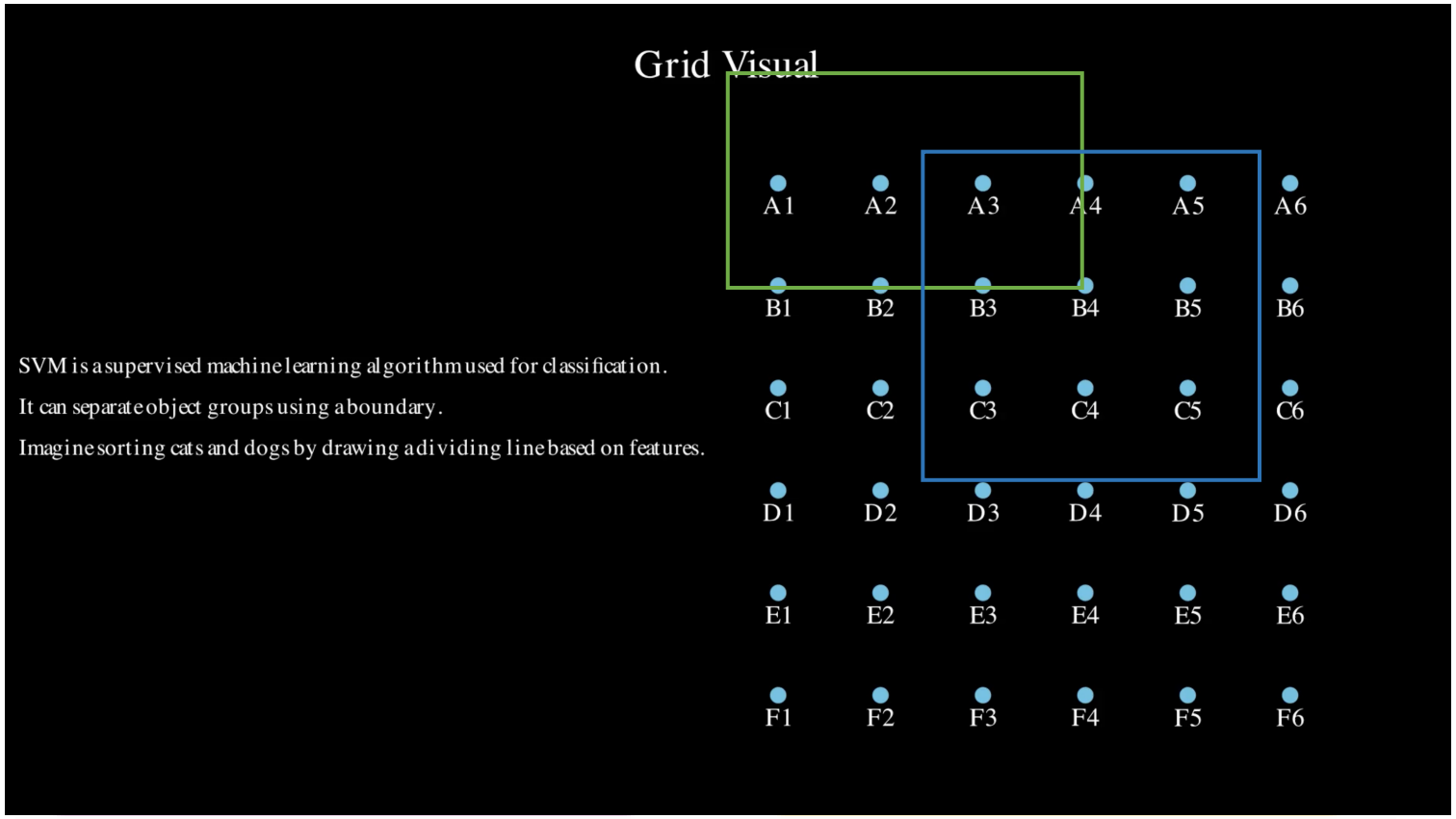


Method (2/3)

- A naive "paste all code + full error log" is expensive as section code may reach hundreds of lines
- ScopeRefine: a hierarchical, scope-guided, Go-to Style debug & fix procedure
  - *Line scope*. Localize to the error line and its immediate context
  - *Lecture line block scope*. Expand to the block bounded by markers for a certain lecture line



Method (3/3)


Grid Visual
A1 A2 A3 A4 A5 A6
B1 B2 B3 B4 B5 B6
C1 C2 C3 C4 C5 C6
D1 D2 D3 D4 D5 D6
E1 E2 E3 E4 E5 E6
F1 F2 F3 F4 F5 F6
SVM is a supervised machine learning algorithm used for classification.
It can separate object groups using a boundary.
Imagine sorting cats and dogs by drawing a dividing line based on features.

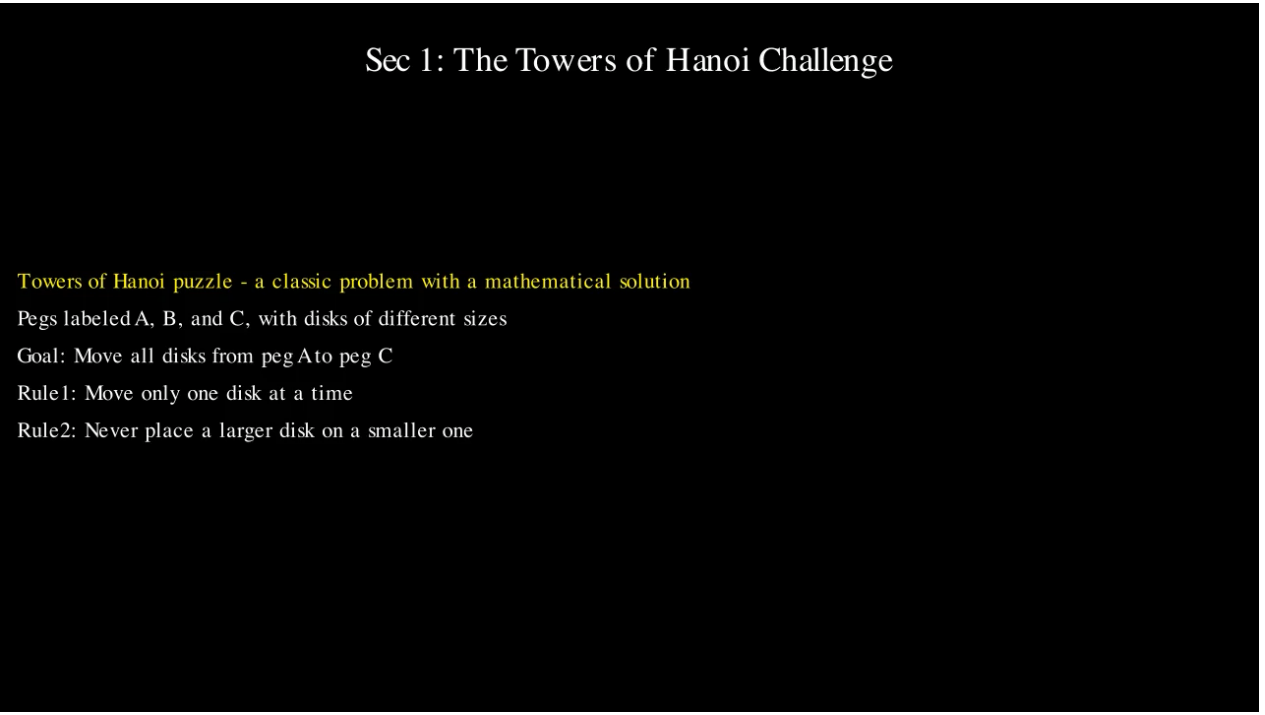
Experiment (1/2)

- *Pixel-based models lag far behind, so does pure CodeLLM. Code2Video makes clear improvements*
- GPT-4o mini runs rapidly, suffers from repeated bug fixes, increase both time & token usage
- GPT-5 tends to concise and efficient code → low cost but underperform in attractiveness
- Claude Opus 4.1 achieves higher AES and TQ scores, highlighting code reliability

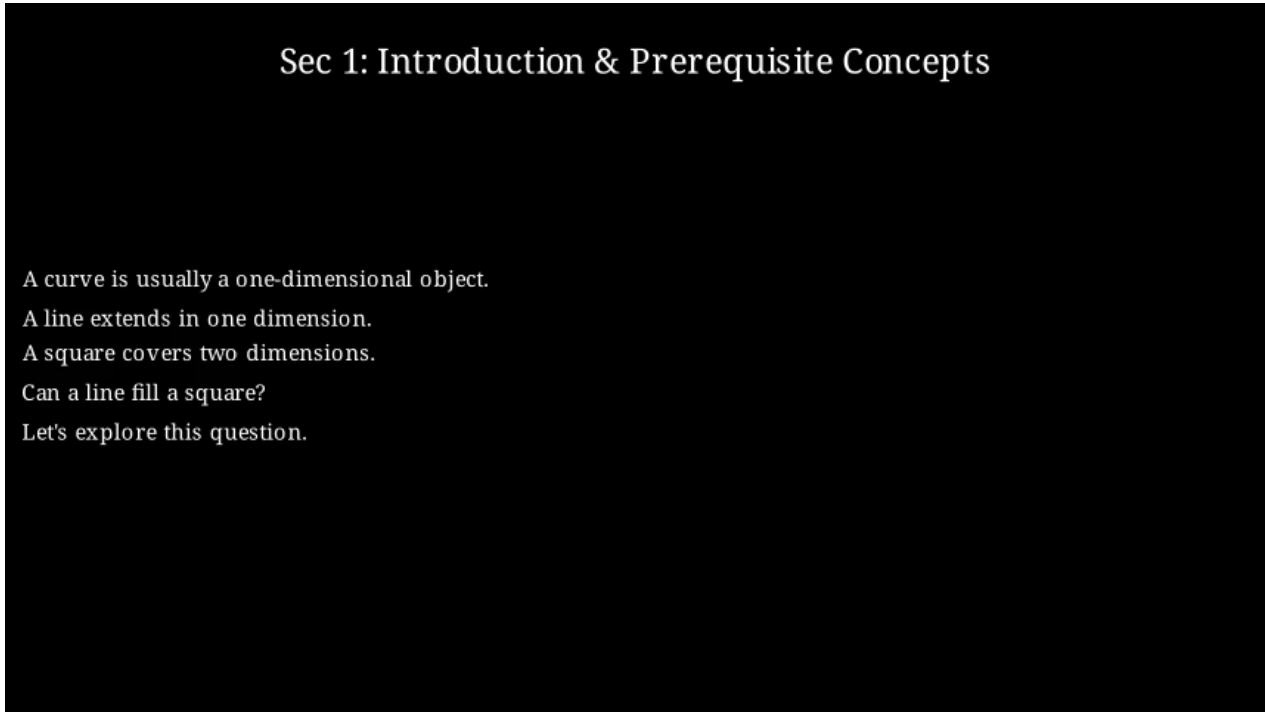
Method	EFF (↓)		AES (↑)						TQ (↑)
	Time	Token (K)	EL	AT	LF	VC	AD	Avg	
Human-made 3B1B	–	–	98.3	100	100	100	100	99.7	44.6
Pixel-based diffusion									
OpenSora-v2	27.6	–	0.0	5.0	0.0	0.0	13.3	3.0	11.0
Wan2.2-T2V-A14B	17.4	–	0.0	10.0	0.0	0.0	20.0	5.0	18.5
Code LLM									
GPT-4.1	2.1	1.2	10.0	15.0	0.0	5.0	30.0	10.8	20.8
Claude Opus 4.1	2.8	2.3	10.0	20.9	11.9	12.0	40.4	17.6	22.1
Code2Video Agent (Ours)									
Code2Video Gemini-2.5 Pro	15.5	41.8	70.3	60.3	44.3	37.6	74.7	55.6	32.7
Code2Video GPT-4o	14.1	32.7	70.3	58.3	54.6	48.5	68.3	59.3	33.5
Code2Video GPT-4o mini	16.8	49.2	77.0	52.8	73.0	57.2	79.0	66.7	34.2
Code2Video GPT-5	8.8	19.3	75.5	60.5	81.8	63.6	79.7	71.4	39.4
Code2Video GPT-4.1	15.4	30.8	82.8	65.6	95.0	68.0	83.7	78.2 +67.4	36.6 +15.8
Code2Video Claude Opus 4.1	13.8	43.1	90.6	79.7	93.3	84.2	91.9	**87.6** +80.0	**40.5** +18.4

Experiment (2/2)

- For more cases: https://chenanno.github.io/Code2Video/



Hanoi Challenge (Puzzles)



Space Filling Curves (Topology)

Future Work

- Exploring how AI can transform education by enabling personalized and scalable teaching for both humans and robots
- Whether instructing humans in symbolic knowledge or robots in physical tasks, long-term, AI-driven guidance remains a key challenge

Thank you!