Video models are zero-shot learners and reasoners

Google Deepmind 2025

Reviewed & Presented by Joon Hyeok Kim

Contents

Recall the history of the language models.

The Pre-LLM Era was like the 群雄割据(군웅할거) of ...

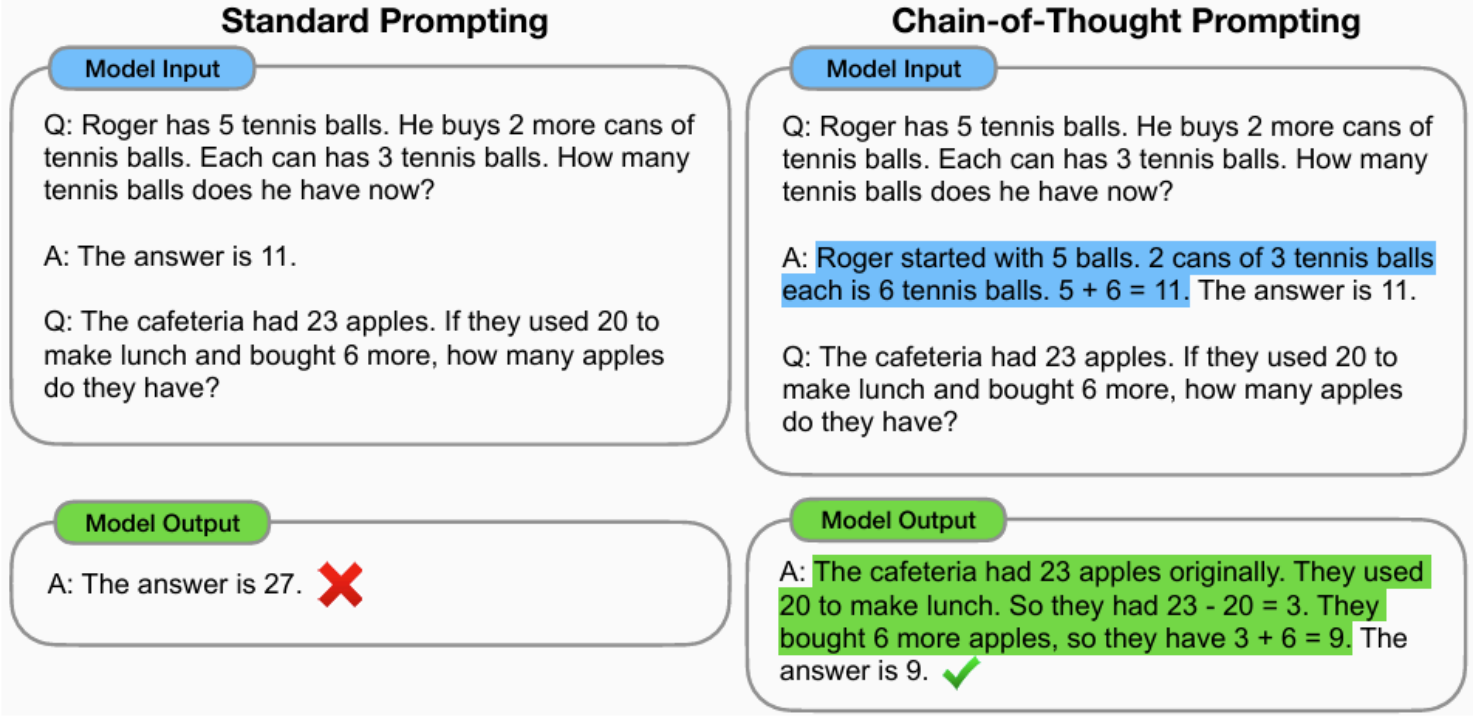
Task-Specific Bespoke Models

- Translation
- Summarization
- Domain specific QnA
- ...

LLM unified and started to work as a Generalist

- Translation
- Summarization
- Domain specific QnA
- And, now even capable of
  - Coding
  - Math
  - Creative writing
  - Deep research (oh my...)
  - and so on...

Paradigm Shift : Chain-of-Thoughts (CoT) + Computing Power

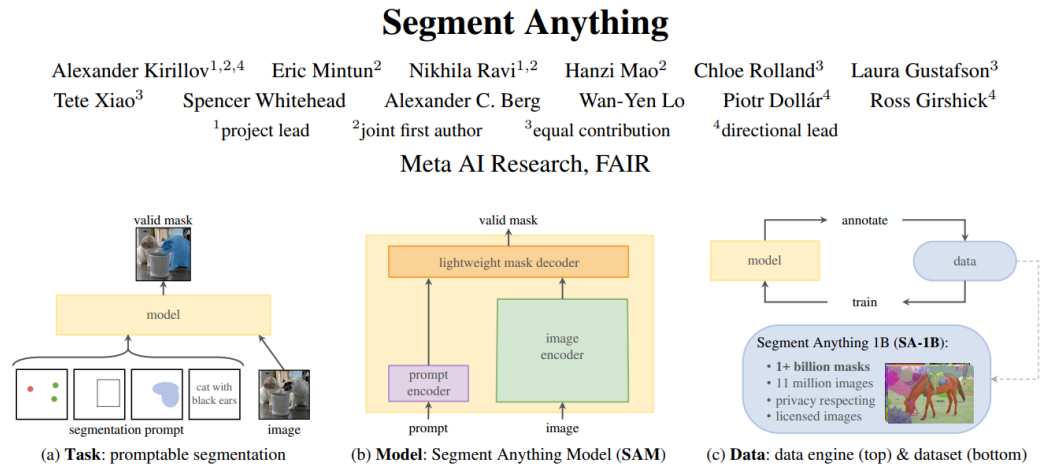


Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
Google Research, Brain Team 2022

Analogy in Video Models

Current task-specific video models that outperform general video models...

Segmentation Specialist



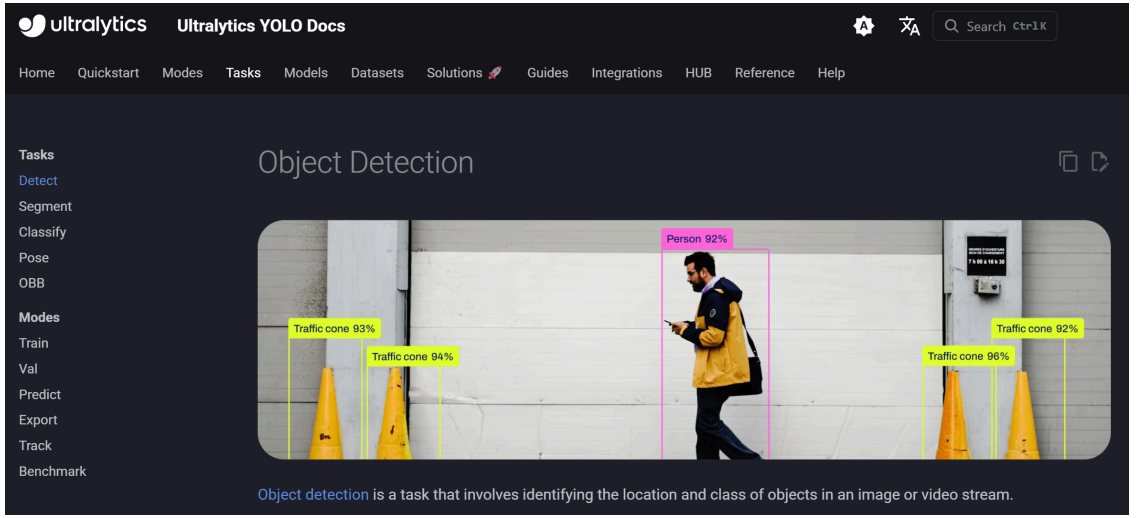
Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR

(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a prompt-able segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

Object Detection Specialist (YOLO variants)



Can video models become Generalists just like LLM did?

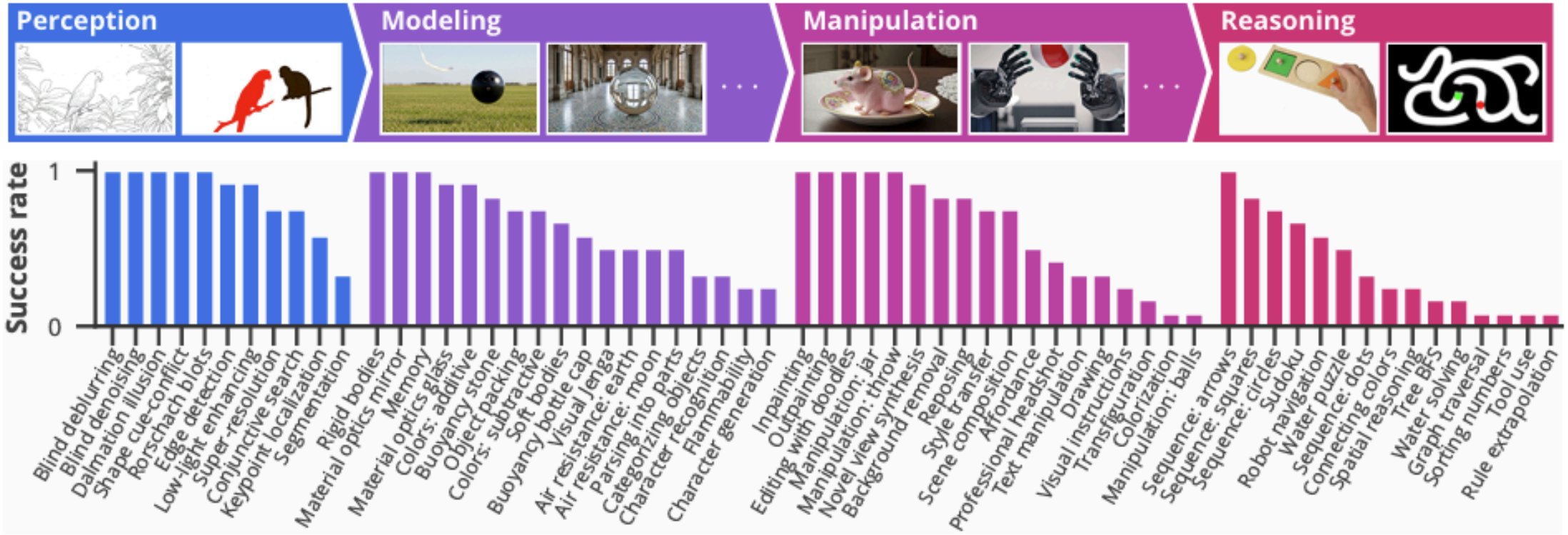
Maybe yes with...

Chain-of-Frames (CoF) + Future Developments

- Applying changes across **dimensions** of the real world frame-by-frame
  - Dimensions : Time & Space
  - Similar to Step-by-step strategy in CoT

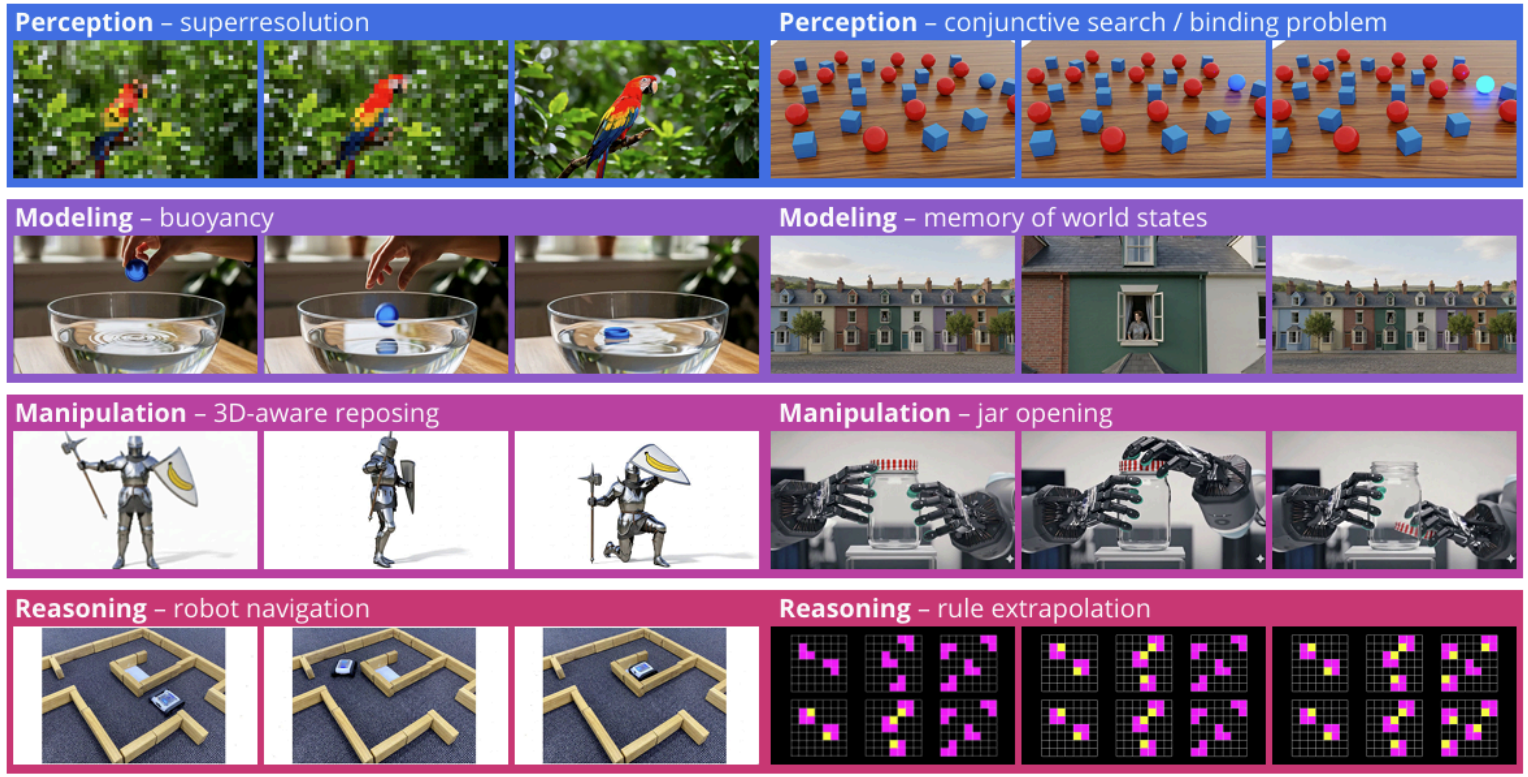
Still, conceptual and no analytic mechanism studied yet...

2. Hierarchical Categorizations of Visual Capabilities



Provides a framework to assess various abilities of Video Models

Stacking up!



Understand the world

Model the visual world

Alter modeled world

Reason across dims.

Examples

3. Methods, Evaluation, & Results

2. Methods

Approach and motivation Our method is simple: We prompt Veo. This minimalist strategy is intentional, as it mirrors the transformation of NLP from task-specific fine-tuning or training to

* *Joint leads.*

e.g.

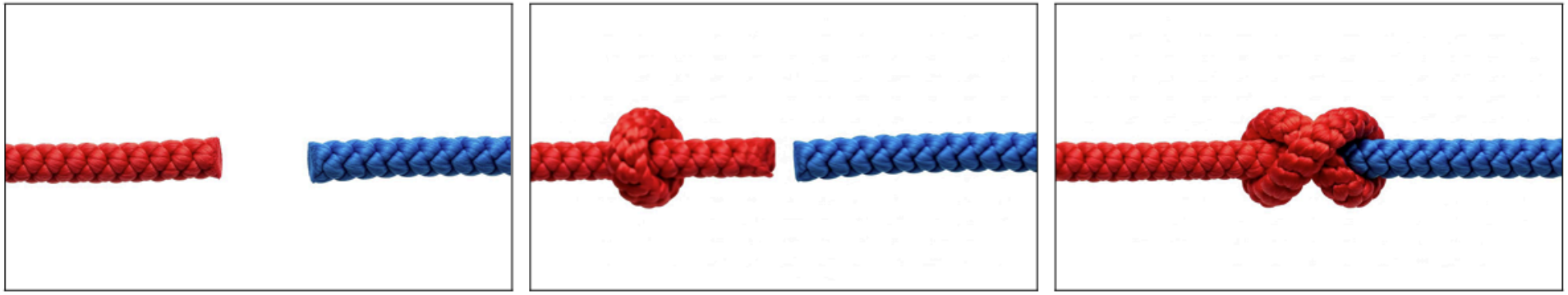
Figure 29 | **Categorizing objects.** Prompt: *"A person puts all the kids toys in the bucket. Static camera, no pan, no zoom, no dolly."* Success rate: 0.33.

Evaluation Methodologies

Qualitative Evaluation

Concept) Success Rate

- Def.)
  - The fraction of generated videos that solved the task
- Props.)
  - Determined by **humans**
  - > 0 : the model possesses the ability to solve the task
  - ≈ 1 : the model reliably solves the problem irrespective of the random seed

Success	 Figure 28 | **Color mixing**. **Additive** (lights, top). Prompt: *"The spotlight on the left changes color to green, and the spotlight on the right changes color to blue."* Success rate: 0.92. **Subtractive** (paints, bottom). Prompt: *"A paintbrush mixes these colors together thoroughly until they blend completely. Static camera, no pan, no zoom."* Success rate: 0.75.
Failure	 Figure 65 | **Tying the knot**. Prompt: *"A knot is tied connecting these two rope ends."* Failure: physics violation, impossible rope movement.

Quantitative Evaluation

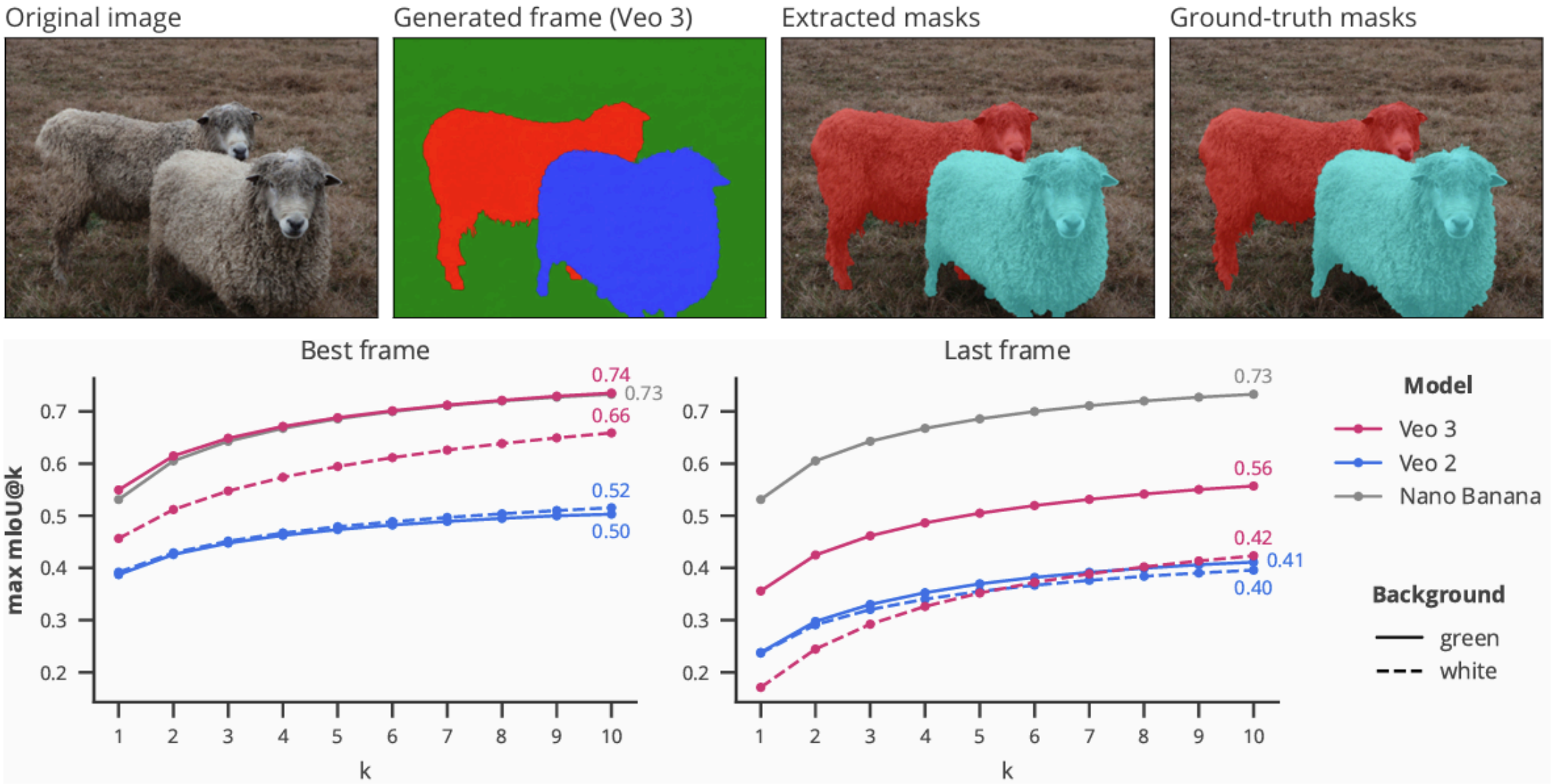
Problem specific scores are used.

- No unified measure nor statistics

Consider both the best frame and the last frame

- Best Frame : Best performance but not deterministic
- Last Frame : Deterministic but no guarantee on performance

e.g.) Segmentation Problem (Perception) : Intersection over Union (IoU)



e.g.) Maze Solving Problem (Reasoning)

Compare...

For Video Models... (Veo 2 & Veo 3)

1. Provide a maze image to the prompt as the first frame

- They tried various maze datasets.
  - `maze-dataset 0.3.4`
  - Hand drawn irregular mazes (flipped/rotated) to get 40 unique samples

2. Prompt text as below

Veo

Create a 2D animation based on the provided image of a maze. The red square slides smoothly along the white path, stopping perfectly on the green square. The red square never slides or crosses into the black areas of the maze. The camera is a static, top-down view showing the entire maze.

Maze:

- *The maze paths are white, the walls are black.*
- *The red square moves to the goal position, represented by a green square.*
- *The red square slides smoothly along the white path.*
- *The red square never slides or crosses into the black areas of the maze.*
- *The red square stops perfectly on the green square.*

Scene:

- *No change in scene composition.*
- *No change in the layout of the maze.*
- *The red square travels along the white path without speeding up or slowing down.*

Camera:

- *Static camera.*
- *No zoom.*
- *No pan.*
- *No glitches, noise, or artifacts.*

Detailed description of the maze problem setting in paragraph form

Iteratively refining the maze description and scenic details in a note-taking format.

For Other Reference Models

- Nano Banana (Image Specific Generative Model)

> **Nano Banana**
> *Mark the correct path from the red to the green circle through the maze in blue.*

- Gemini 2.5 Pro (Language Model)

> **Gemini 2.5 Pro I2T**
> SYSTEM
> *Think step by step as needed and output in xml format:*
> *<think>thinking process</think>*
> *<final_answer>final answer</final_answer>*
>
> *USER*
> *The following image shows a maze, represented by colored squares:*
>
> - *Black squares represent walls and cannot be passed through.*
> - *White squares are empty and can be passed through.*
> - *The red square is the starting point.*
> - *The green square is the end point.*

Evaluation

Define Illegal Moves

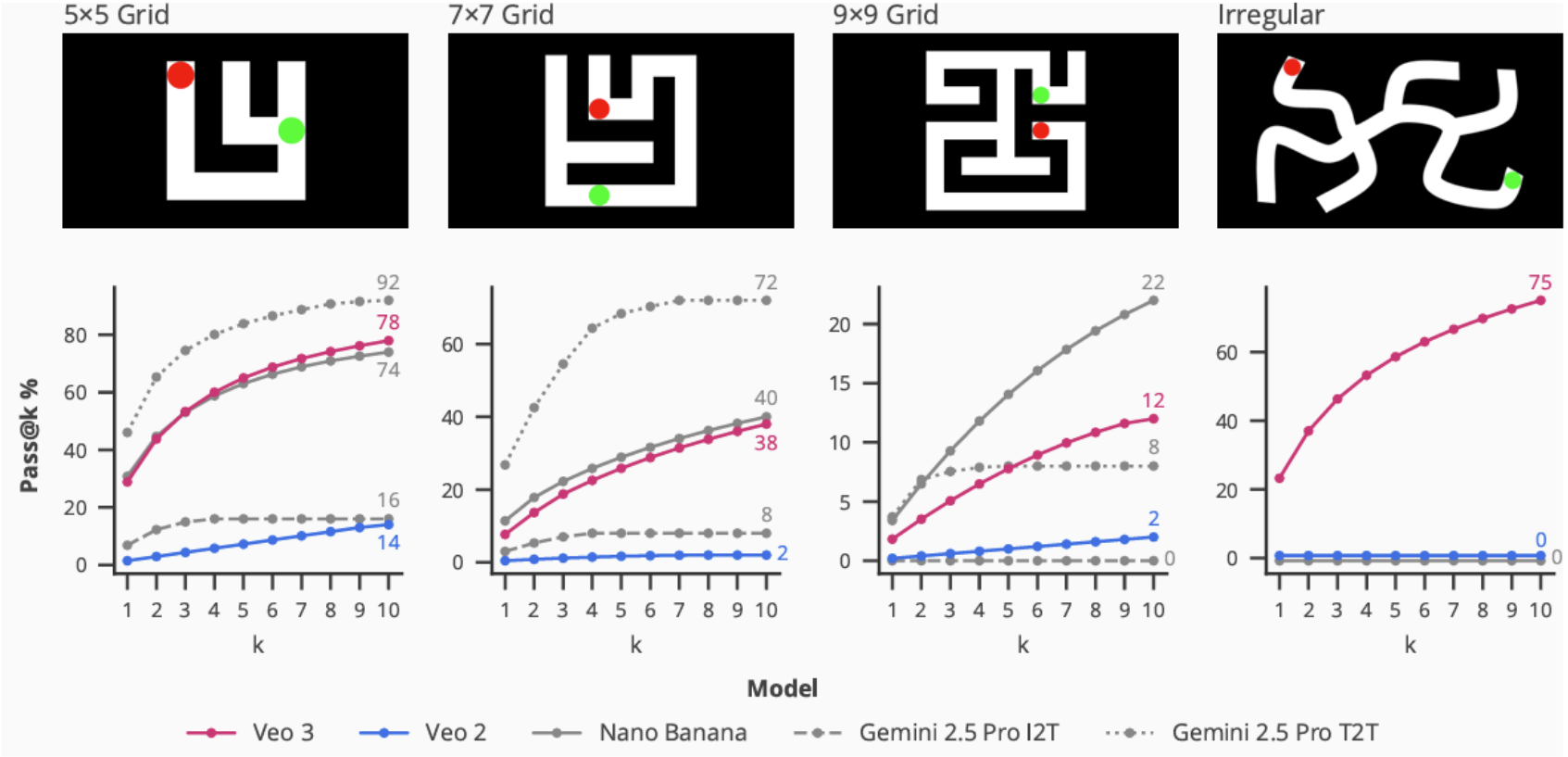
- Jumping over walls
- Clipping through boundaries
- Alteration of the goal's position
- ...

Success Rate (Quantitative)

- The fraction of k attempts where the agent successfully reaches the goal without illegal moves through out the generated video.

Result

- Veo2 vs Veo3
- $k \uparrow \Rightarrow \text{pass} \uparrow$
- Complex
  - Beats LLM
- Simple, Irregular
  - Beats Image



Limit : Insufficient Reasoning Capability

Significant portion of the reasoning experiments scored low success rates.

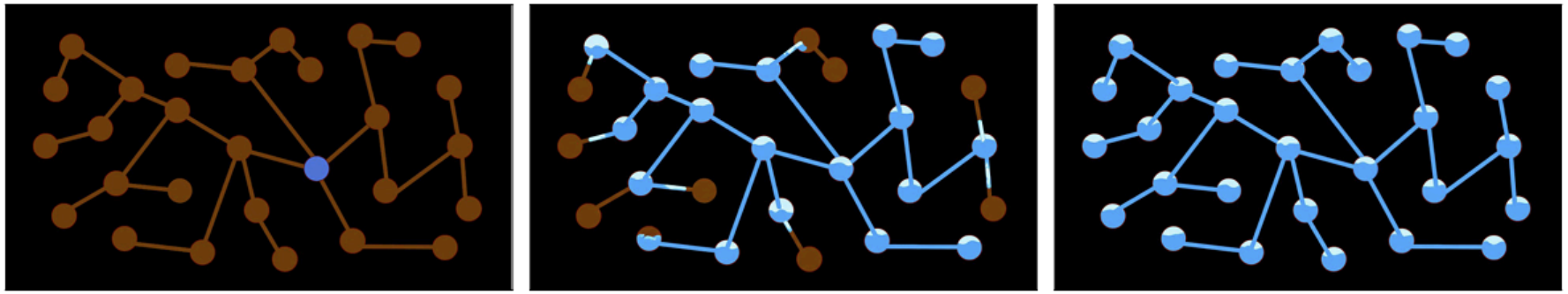


Figure 48 | **Graph traversal**. Prompt: *"Starting from the blue well, an unlimited supply of blue water moves through the connected channel system without spilling into the black area."* Success rate: 0.08.

Figure 57 | **Maze solving**. Prompt: *"Without crossing any black boundary, the grey mouse from the corner skillfully navigates the maze by walking around until it finds the yellow cheese."* Success rate: 0.17.

Failures on complex tasks

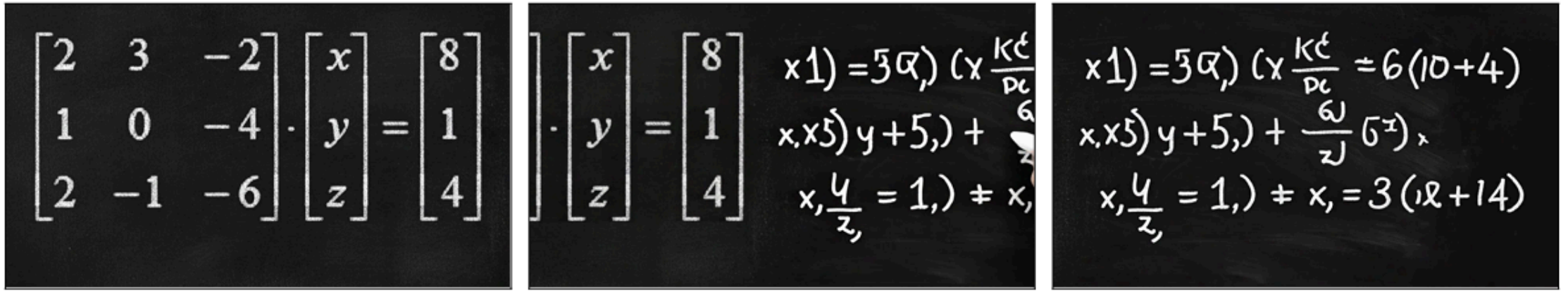


Figure 69 | **Solving system of linear equation.**. Prompt: *"A hand appears and solves the set of linear equations. It replaces the x, y, z matrix with their correct values that solves the equation. Do not change anything else."* Failure: hallucinations with text on the blackboard.

Figure 72 | **Glass falling.** Prompt: *"The object falls. Static camera, no pan, no zoom, no dolly."* Failure: physics violation, glass does not break, and orients itself to be vertical after landing on the floor.

More Limits...

No analytic relation proven between CoF and Zero-shot learning

- Does CoF guarantees the correct reasoning path?
- Or, does "correctness" even matter if we have the answer?

No unified evaluation method : Human and problem dependent

Video generation is still too computationally expensive

- Emergent abilities only at scale! (Veo 3's performance improvement)

High Dependency on Prompt Engineering

- Nevertheless, CoT did change the world...

Jack of many trades, master of few

- Fine-tuned models dominate in specific tasks. Will they last forever?

Future Outlook

Video Models Show Great Potential for Zero-Shot Learning

- Distinguish a model's task performance and its underlying ability to solve it.
- Early LLMs underperformed against fine-tuned models.
  - But look who dominates now, huh?

We are at the very beginning of the Video Model development

- Improvement from Veo 2 to Veo 3 indicates further developments.
- Alternative approach on Prompt Engineering?
  - Current : first frame image + text description
  - Just like CoT changed the game with the prompt engineering in LLMs.
- Cost will fall : LLM's ongoing decrease in cost may support this.

Leveraging Scaling Laws and Optimization

- Performance can be further boosted by applying inference-time scaling and standard optimization toolkits, which were not used in this study

Questions

Thank you