Project Part 3

Computational Visual Perception (CompVP)

Bernhard Egger, Andreas Kist, Patrick Krauß, Tim Weyrich

Overall project goal

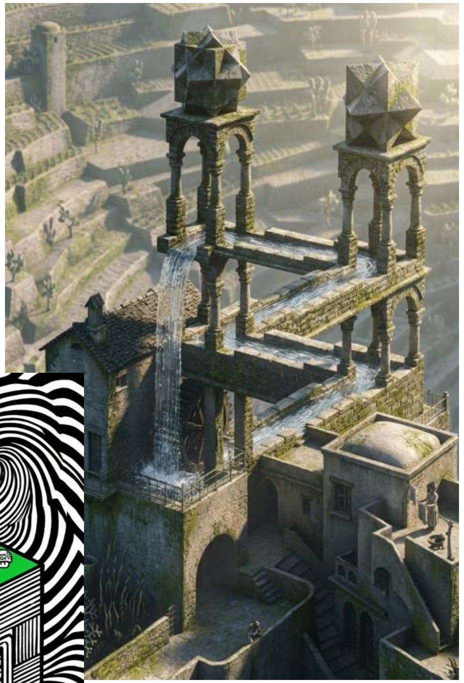
- How much of vision do SotA generative video models “solve”?

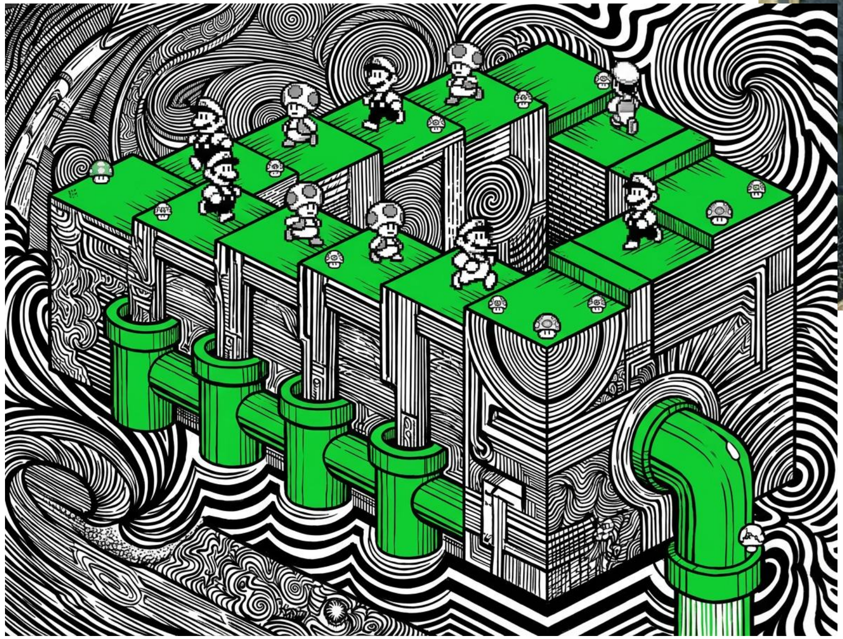
- How well do they and other models work for corner cases of vision?

Part 2 results and grading

- Common pitfalls
- Favorites & Recommended

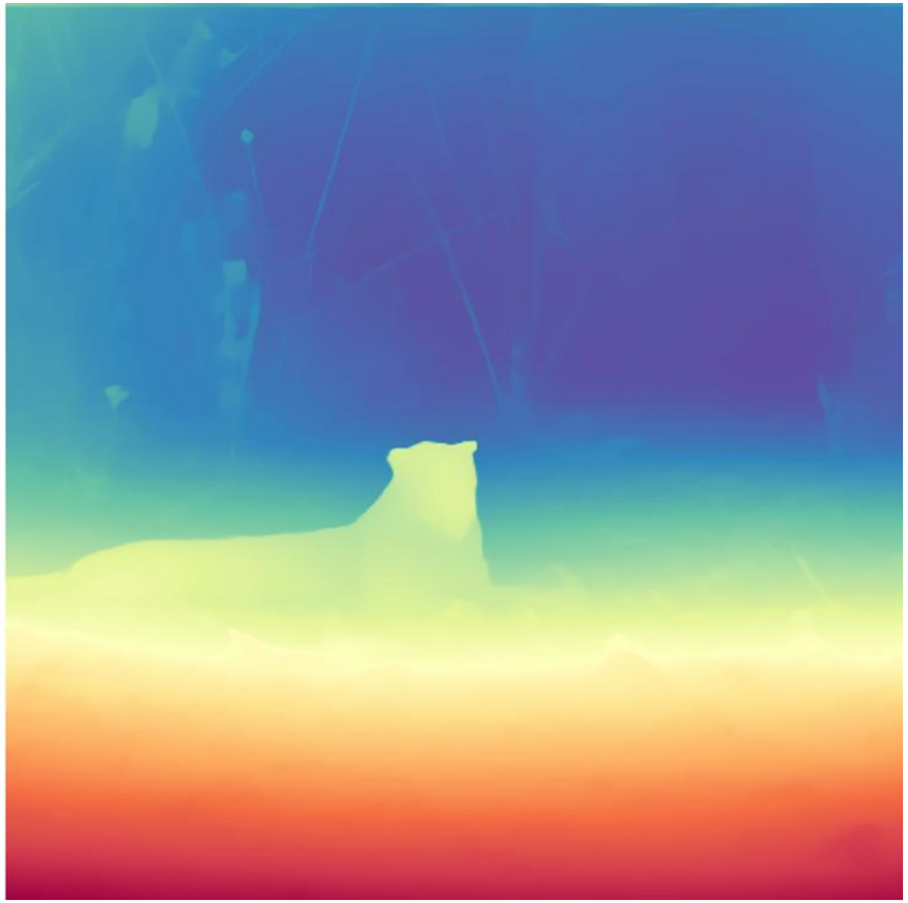
Part 2 my favorites: Berger Correa Tran



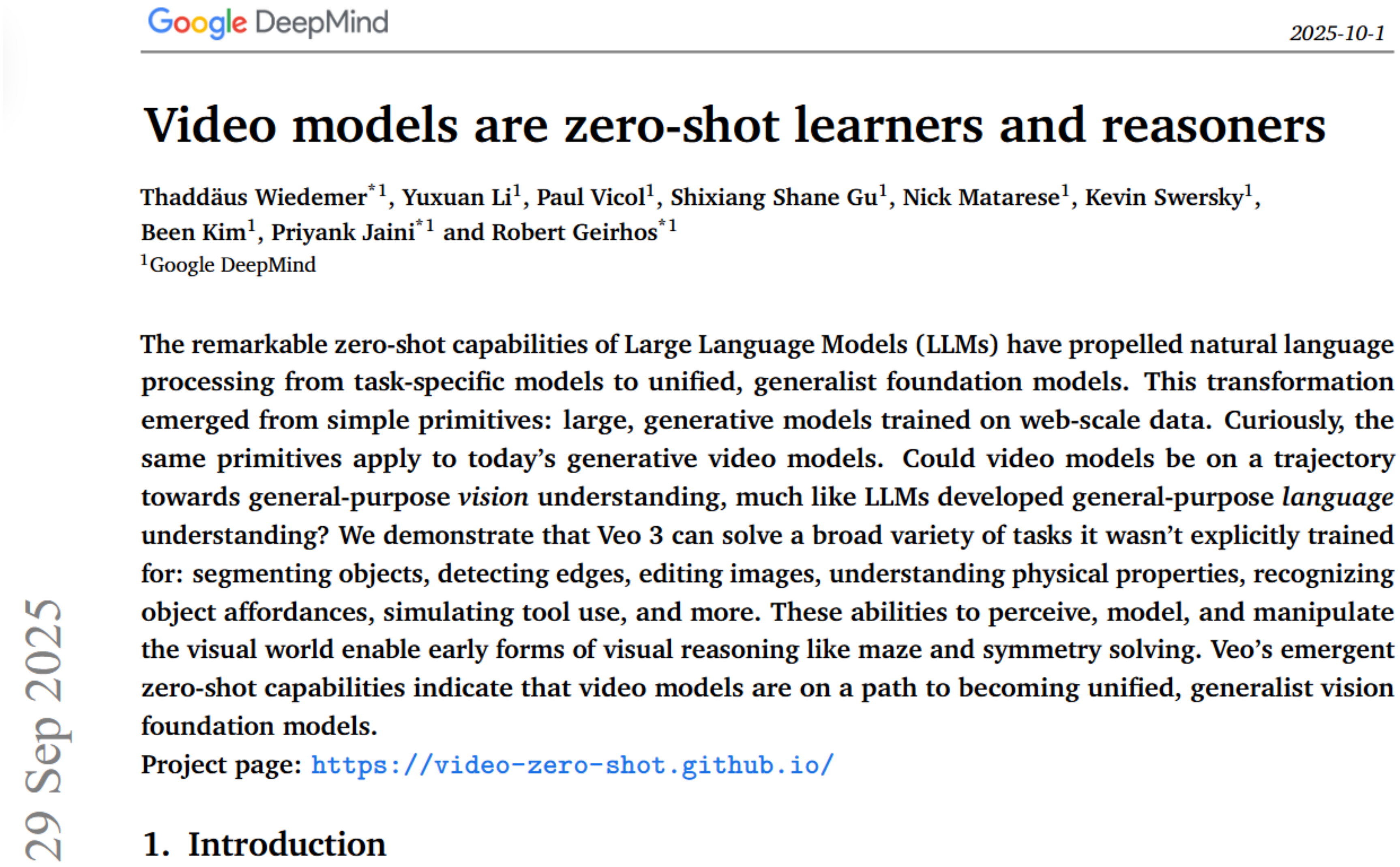


Part 2 my favorites: Nasirkhani

Part 2 my favorites: Kamrishi



Paper 1:



Google DeepMind

2025-10-1

Video models are zero-shot learners and reasoners

Thaddäus Wiedemer*[1], Yuxuan Li[1], Paul Vicol[1], Shixiang Shane Gu[1], Nick Matarese[1], Kevin Swersky[1], Been Kim[1], Priyank Jaini*[1] and Robert Geirhos*[1]

[1] Google DeepMind

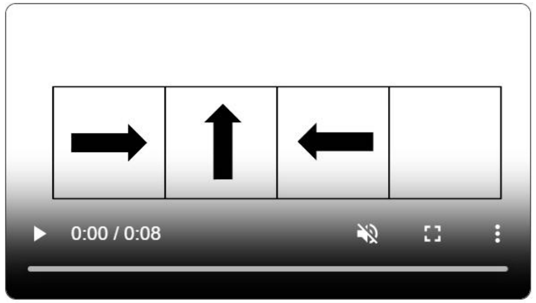
The remarkable zero-shot capabilities of Large Language Models (LLMs) have propelled natural language processing from task-specific models to unified, generalist foundation models. This transformation emerged from simple primitives: large, generative models trained on web-scale data. Curiously, the same primitives apply to today's generative video models. Could video models be on a trajectory towards general-purpose *vision* understanding, much like LLMs developed general-purpose *language* understanding? We demonstrate that Veo 3 can solve a broad variety of tasks it wasn't explicitly trained for: segmenting objects, detecting edges, editing images, understanding physical properties, recognizing object affordances, simulating tool use, and more. These abilities to perceive, model, and manipulate the visual world enable early forms of visual reasoning like maze and symmetry solving. Veo's emergent zero-shot capabilities indicate that video models are on a path to becoming unified, generalist vision foundation models.

Project page: https://video-zero-shot.github.io/

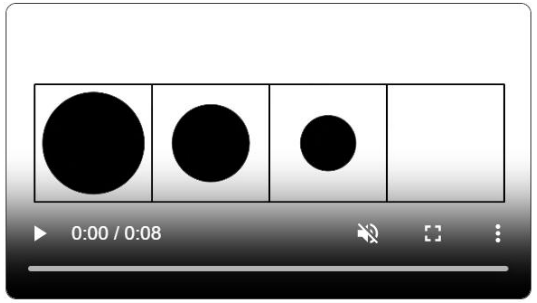
29 Sep 2025

1. Introduction

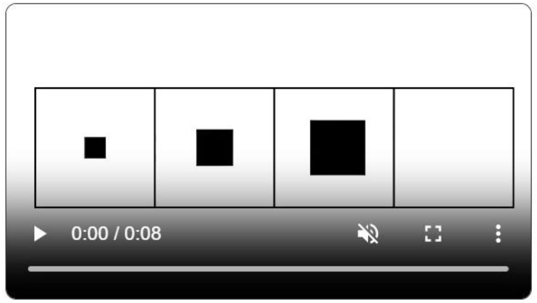
Paper 1 Reasoning:



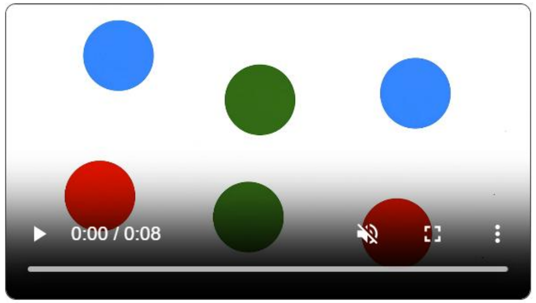
Sequence (arrows)



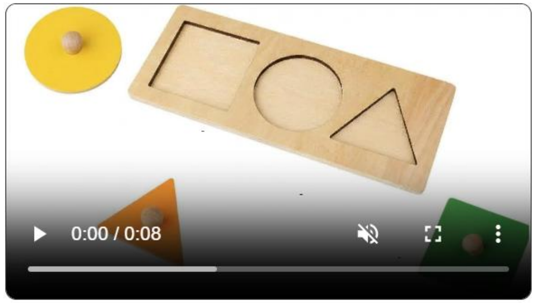
Sequence (circles)



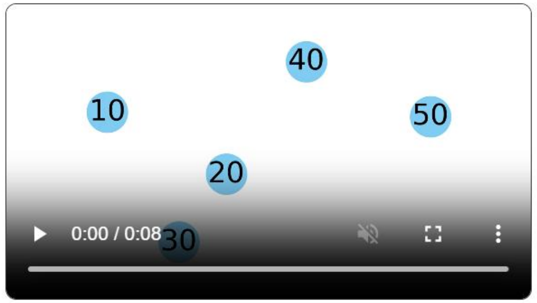
Sequence (squares)



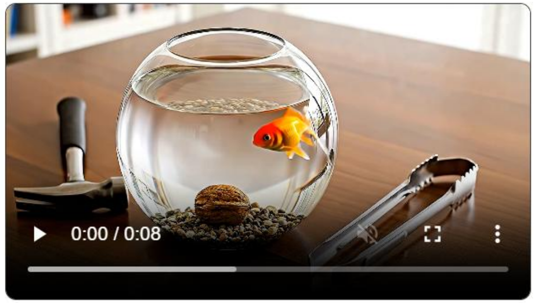
Connecting colors



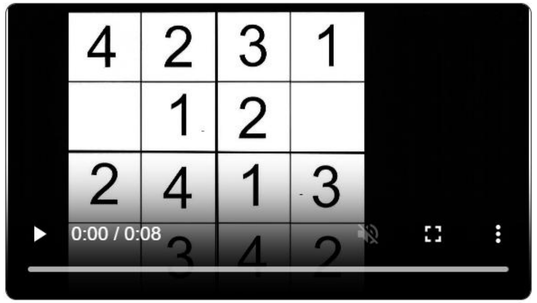
Shape fitting



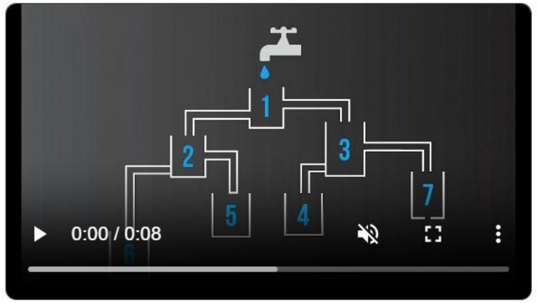
Sorting numbers



Tool use

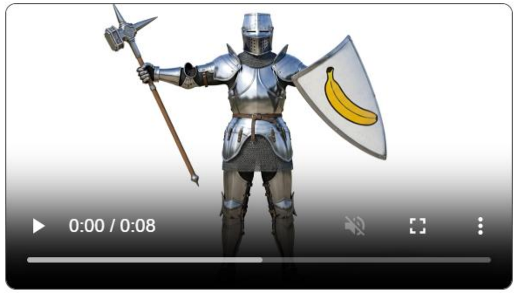


Simple sudoku completion

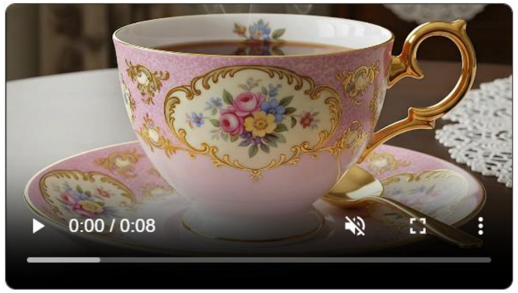


Water puzzle solving

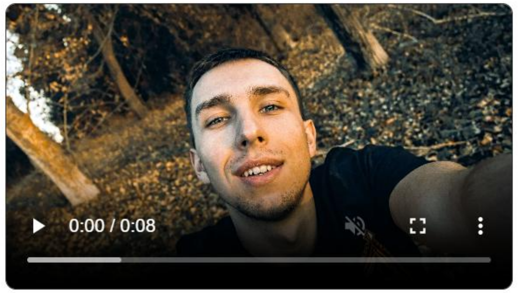
Paper 1 Manipulation:



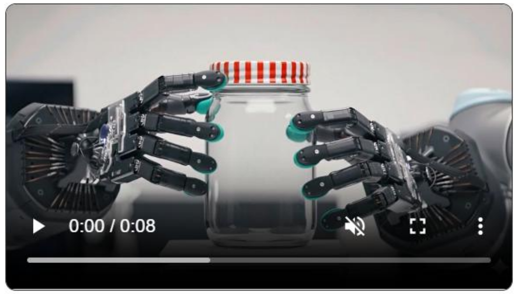
3D-aware reposing



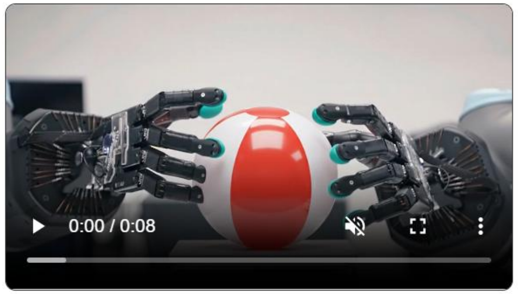
Transfiguration



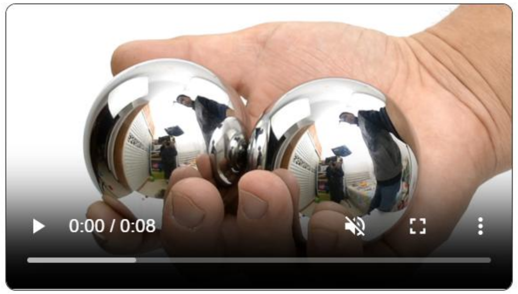
Professional headshot



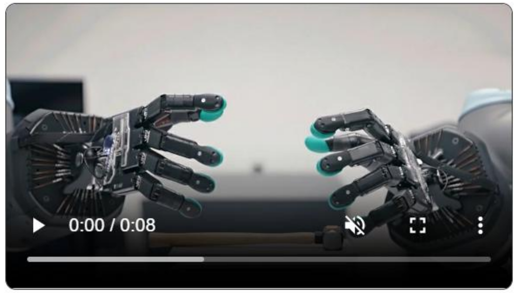
Dexterous manipulation (jar)



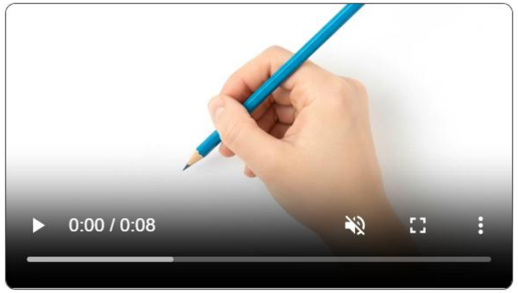
Dexterous manipulation (throw/catch)



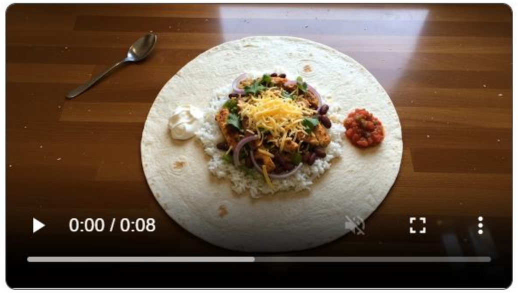
Dexterous manipulation (baoding balls)



Affordance recognition



Drawing



Visual instruction (burrito rolling)

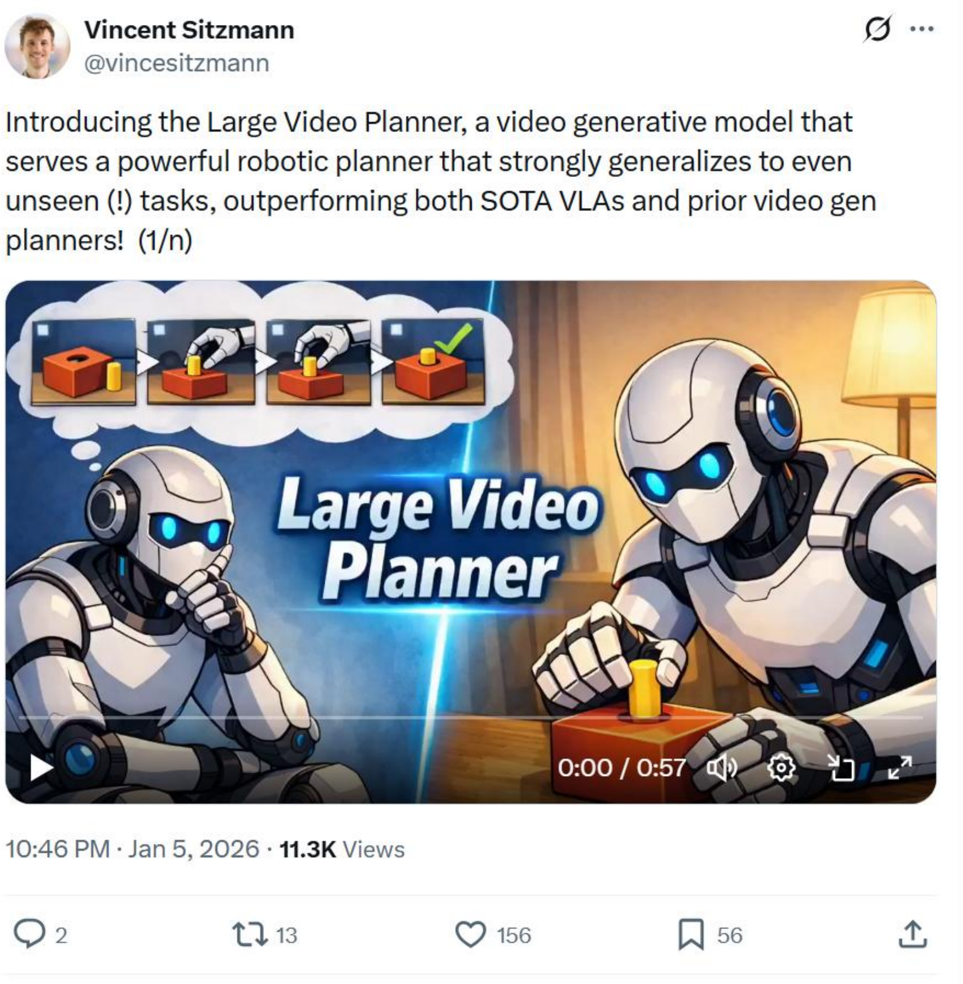
Structure of project

- Project Part 1: Play around with video diffusion models and reproduce paper
  (close to Paper 1)

- Project Part 2: Experiment further with illusions, optical flow and depth estimation
  (close to paper 2)

- Project Part 3: Further Evaluation based on results of Part 1 and Part 2
  (close to paper 1 and 2)

What happened since?



Vincent Sitzmann
@vincesitzmann

Introducing the Large Video Planner, a video generative model that serves a powerful robotic planner that strongly generalizes to even unseen (!) tasks, outperforming both SOTA VLAs and prior video gen planners! (1/n)

10:46 PM · Jan 5, 2026 · **11.3K** Views

What happened since?

Large Video Planner Enables Generalizable Robot Control

oyuan Chen*, Tianyuan Zhang*, Haoran Geng*, Kiwhan Song, Caiyi Zha
Peihao Li, Jitendra Malik, Pieter Abbeel, William T. Freeman,
Russ Tedrake, Vincent Sitzmann, Yilun Du

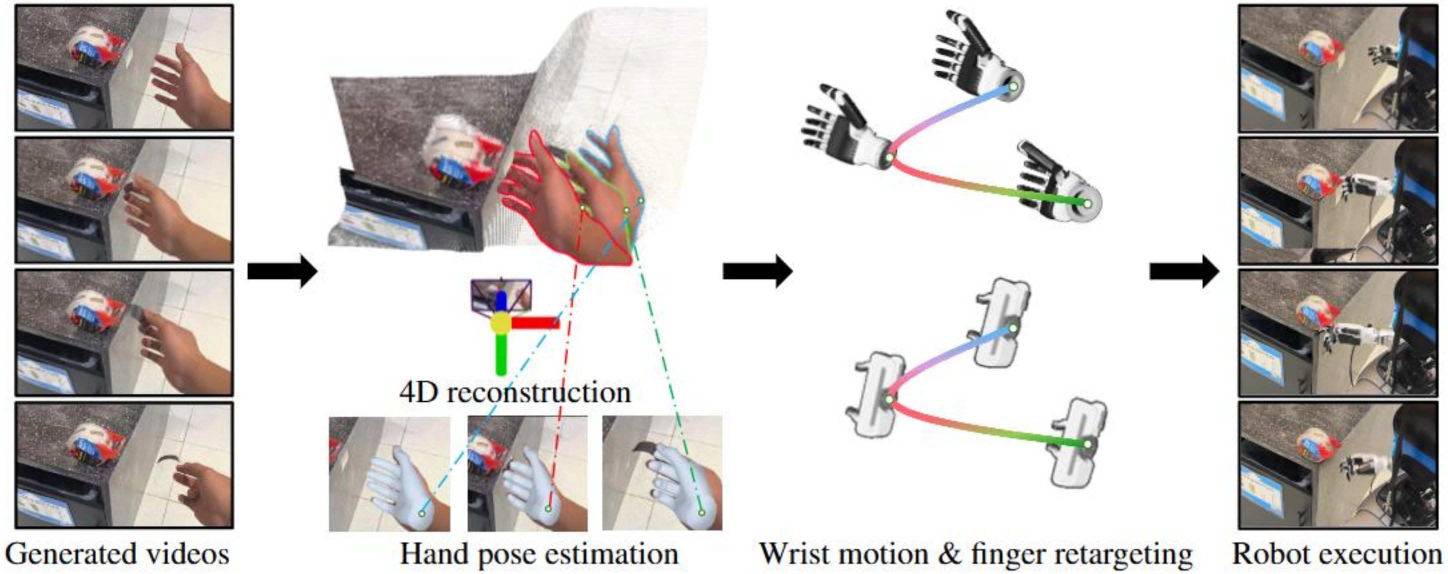
Then retarget to real robot

Part 3

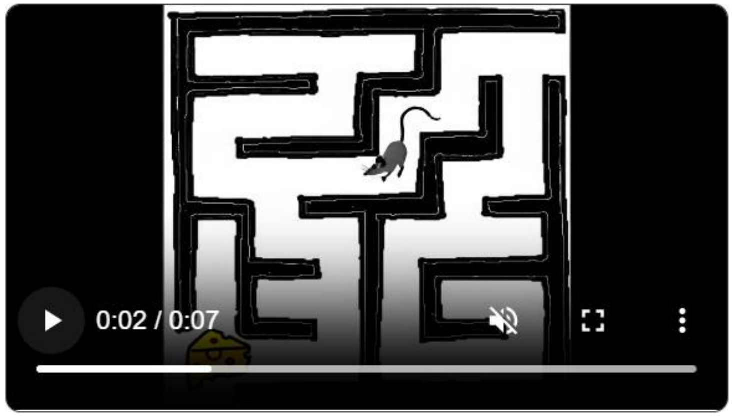
- Choose a Task a Video Model can solve

- Extract Solution

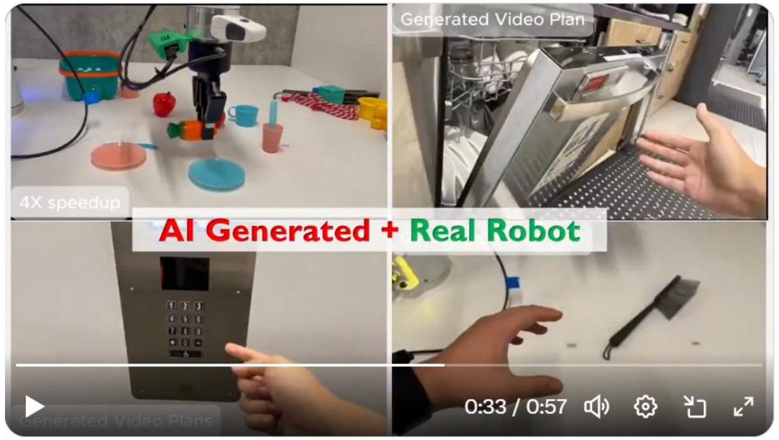
- ~~Retarget~~

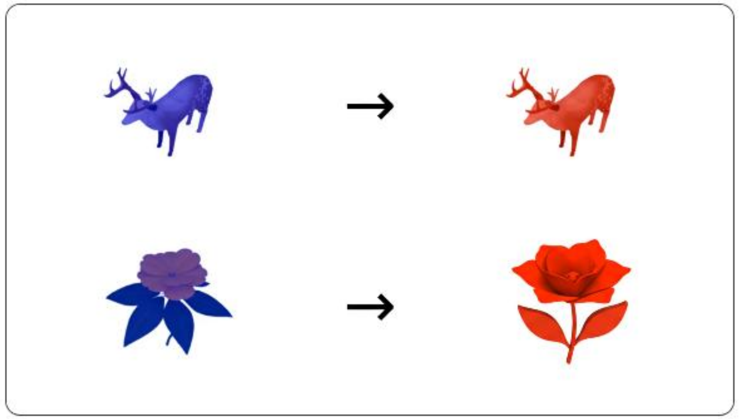


Part 3 Example Tasks



Maze solving (mouse)





Analogy (color)

. . .

How to pass

- Submission via Studon course:
  “Computational Visual Perception”

- Submission has to be in the exact format

- Three strict project deadlines
  - November 21st,
  - January 2nd (feel free to submit early),
  - February 5th

How to pass

- Pass/Fail for each of the 3 parts,
- You need to pass 2 out of 3 parts

- The best solutions for each part will be released ~ 1 week after the deadline, to enable others to continue with the best solution of another team

How to pass

- Scope of project ~ 150 hours per student
- Teams of 1-3 students
- Steps can be performed in new group

- If you are looking for a group, please stay after the class and talk to people who also stay
- If you are looking for a group and can only join virtually, please use the forum in “Computational Visual Perception” to team up

- Finding a group is your responsibility

Part 3

- Choose one task
  - Come up with your own strategy for extraction of a solution from a Video Model
  - Automate it Run it on your own (not just cloud-ui)

Part 3 deliverables

- Per project team 1 single pptx file (or powerpoint compatible)
- The pptx file contains:
  - Explanation of Task
  - Example Videos
  - Explanation of Extraction Strategy
  - Example Extraction
  - Conclusion
  - Screenshot how your interface works (e.g. terminal)

Part 3 Upload

- In Studon course :
  “Computational Visual Perception”
- You will upload up to ~1GB (don’t!)
  Plan in internet speed, upload at university

  If you run into issues uploading, you send an md5 hash of your zip file **before** the deadline and you provide an alternative download link within 24h

Part 3 grading

- Explanation of Task
- Example Videos
- Explanation of Extraction Strategy
- Example Extraction
- Conclusion
- Screenshot how your interface works (e.g. terminal)

- Selected solution: all points fulfilled to full satisfaction
- Pass: **no bullet point from the above missing**

- Plagiarism will have serious consequences

Project Consulting

- You can ask questions in the forum "Computational Visual Perception"

- You come with concrete questions
- I'll open a thread in the forum, where you can respond till Thursday each week if you want to meet
- I'll distribute time slots each Friday
- Grade prediction possible in project consulting session

- No guarantee for any responses on the day of the deadline

How to complete module

- 7,5 ECTS
  - 5+2,5 ECTS
  - You need both!
  - You can’t get only 5 or only 2,5 ECTS

Don’t start late