Friedrich-Alexander-Universität
Technische Fakultät

FAU

Project Part 1

Computational Visual Perception (CompVP)

Bernhard Egger, Andreas Kist, Patrick Krauß, Tim Weyrich

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

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?

Paper 1 by October 27th :

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/

1. Introduction

Paper 2 by November 27th :

frontiers in Psychology

ORIGINAL RESEARCH
published: 15 March 2018
doi: 10.3389/fpsyg.2018.00345

Check for updates

Illusory Motion Reproduced by Deep Neural Networks Trained for Prediction

Eiji Watanabe[1,2], Akiyoshi Kitaoka[3], Kiwako Sakamoto[4,5], Masaki Yasugi[1] and Kenta Tanaka[6]*

[1] Laboratory of Neurophysiology, National Institute for Basic Biology, Okazaki, Japan, [2] Department of Basic Biology, The Graduate University for Advanced Studies (SOKENDAI), Miura, Japan, [3] Department of Psychology, Ritsumeikan University, Kyoto, Japan, [4] Department of Physiological Sciences, The Graduate University for Advanced Studies (SOKENDAI), Miura, Japan, [5] Division of Integrative Physiology, National Institute for Physiological Sciences (NIPS), Okazaki, Japan, [6] Sakura Research Office, Wako, Japan

The cerebral cortex predicts visual motion to adapt human behavior to surrounding objects moving in real time. Although the underlying mechanisms are still unknown, predictive coding is one of the leading theories. Predictive coding assumes that the brain's internal models (which are acquired through learning) predict the visual world at all times and that errors between the prediction and the actual sensory input further refine

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)

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

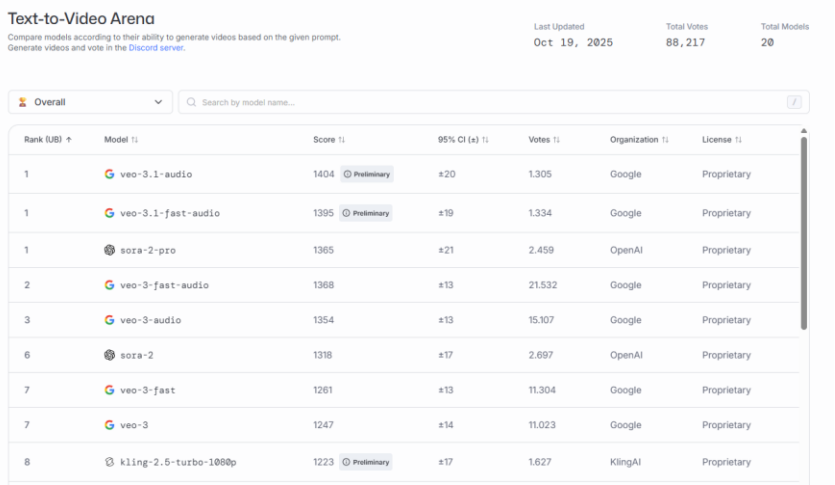
- Scope of project ~ 150 hours per student
- Teams of 1-3 students
- Steps can be performed in new group

- 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 1 Task 1

- Choose 3 cases you would like to reproduce
- Reproduce those cases with the model of your choice
  - Model examples:
    - Veo 3 (closed)
    - Wan 2.2 (open)
    - https://lmarena.ai/leaderboard/text-to-video



Text-to-Video Arena

Compare models according to their ability to generate videos based on the given prompt.
Generate videos and vote in the Discord server.

Last Updated: Oct 19, 2025 | Total Votes: 88,217 | Total Models: 20

Overall | Search by model name...

Rank (UB)	Model	Score	95% CI (±)	Votes	Organization	License
1	veo-3.1-audio	1404 Preliminary	±20	1,305	Google	Proprietary
1	veo-3.1-fast-audio	1395 Preliminary	±19	1,334	Google	Proprietary
1	sora-2-pro	1365	±21	2,459	OpenAI	Proprietary
2	veo-3-fast-audio	1368	±13	21,532	Google	Proprietary
3	veo-3-audio	1354	±13	15,107	Google	Proprietary
6	sora-2	1318	±17	2,697	OpenAI	Proprietary
7	veo-3-fast	1261	±13	11,304	Google	Proprietary
7	veo-3	1247	±14	11,023	Google	Proprietary
8	kling-2.5-turbo-1080p	1223 Preliminary	±17	1,627	KlingAI	Proprietary

Part 1 Task 2

- Make your own variant for each of the 3 cases you did choose
- Take your own image with a camera
  - No AI generated image
  - No images from the internet

Part 1 Task 2 voluntary part

- Collect ground truth (if applicable)

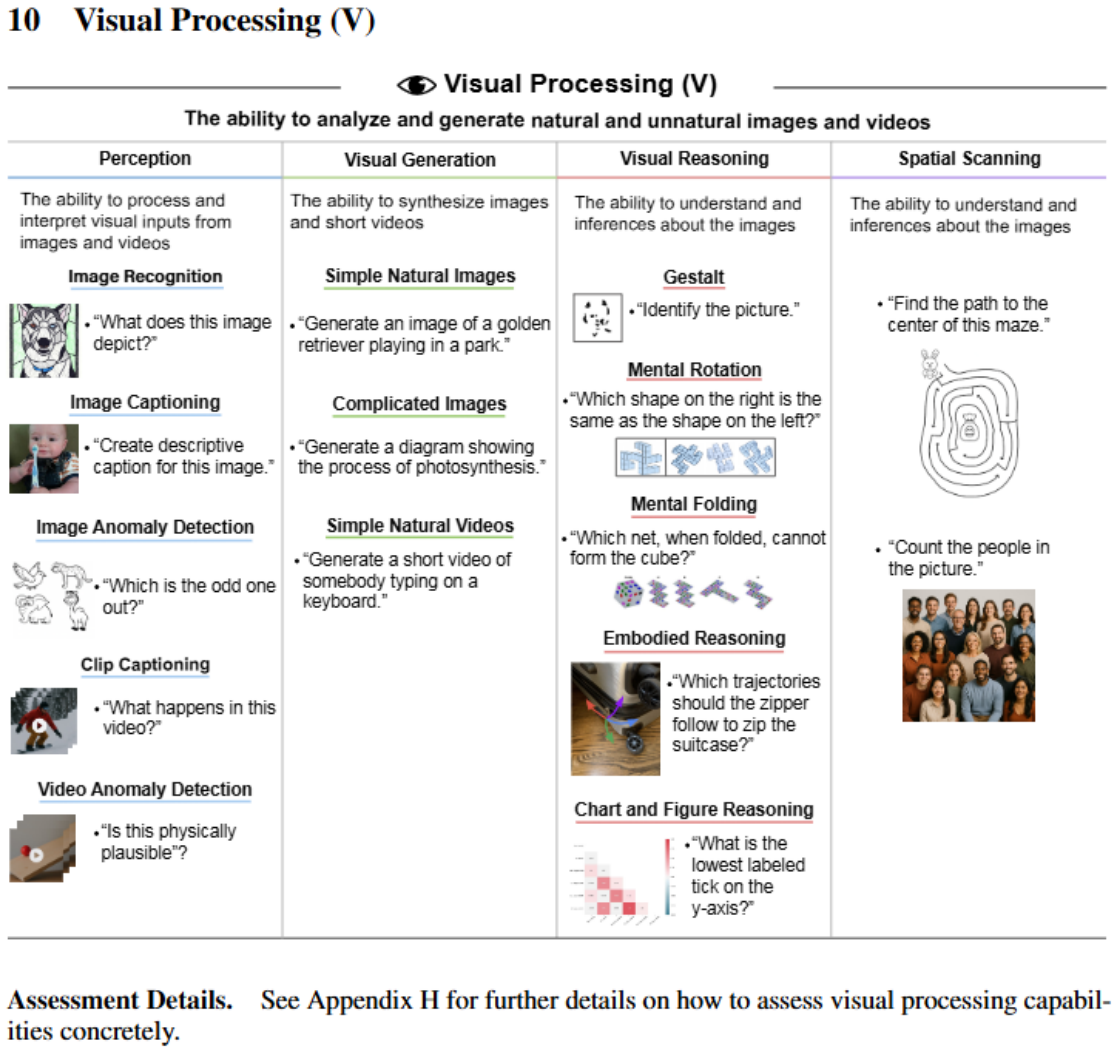
Part 1 Task 3

- Come up with one own task
- If you need inspiration:

A Definition of AGI

Dan Hendrycks[1], Dawn Song[2], Christian Szegedy[3], Honglak Lee[4,5], Yarin Gal[6], Erik Brynjolfsson[7], Sharon Li[8], Andy Zou[1,9,10], Lionel Levine[11], Bo Han[12], Jie Fu[13], Ziwei Liu[14], Jinwoo Shin[15], Kimin Lee[15], Mantas Mazeika[1], Long Phan[1], George Ingebretsen[1], Adam Khoja[1], Cihang Xie[16], Olawale Salaudeen[17], Matthias Hein[18], Kevin Zhao[19], Alexander Pan[2], David Duvenaud[20,21], Bo Li[22], Steve Omohundro[23], Gabriel Alfour[24], Max Tegmark[17], Kevin McGrew[25], Gary Marcus[26], Jaan Tallinn[27], Eric Schmidt[17], Yoshua Bengio[28,29]

[1]Center for AI Safety [2]University of California, Berkeley [3]Morph Labs
[4]University of Michigan [5]LG AI Research [6]University of Oxford [7]Stanford University
[8]University of Wisconsin–Madison [9]Gray Swan AI [10]Carnegie Mellon University
[11]Cornell University [12]Hong Kong Baptist University [13]HKUST
[14]Nanyang Technological University [15]KAIST [16]University of California, Santa Cruz
[17]Massachusetts Institute of Technology [18]University of Tübingen [19]University of Washington
[20]University of Toronto [21]Vector Institute [22]University of Chicago
[23]Beneficial AI Research [24]Conjecture [25]Institute for Applied Psychometrics
[26]New York University [27]CSER [28]Université de Montréal [29]LawZero



Assessment Details. See Appendix H for further details on how to assess visual processing capabilities concretely.

- Or from the lecture

Part 1 Task 4

- Retry everything with a different model
  - Older/newer version
  - Open Source vs. commercial
  - …

Part 1 Task 5

- In one of the models you tried: deactivate prompt rewriting
  - Wan 2.2. definitely possible
  - Veo over API

  - (that needs to be taken into account when choosing models)

Part 1 Task 6

- Summarize results in form of a presentation
- (instead of a report, there will be no actual presentation)

Part 1 deliverables

- Per project team 1 single pptx file (or powerpoint compatible)
- The pptx file contains:
  - Title slide with all team members names
  - 3x1 slide for each reproduced result
    - Titel: used model
    - Left original video, right reproduced video
    - Bottom prompts
  - 3x1 slide for your own variant of reproduced task
    - Titel: used model
    - Left original input image and your input image
    - Right generated videos
    - Bottom prompts
  - Generate one task yourself, same format
  - 2 x 3x1 +1 slide: All results above with a different model, same format
  - 2 x 3x1 +1 slide: All results above also without prompt rewriting, same format
  - 1 slide summarizing what you observed (text)
  - Failure cases (if any)
- All videos must be embedded in the file (no external links)
- (ideally all videos are set to play automatically)

Part 1 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 1 grading

- Correct format
- 3 results reproduced
- 3 own variants videos
- Own task
- All 7 videos above with a different model
- All above 7 videos without prompt rewriting
- Summary text

- Selected solution: all points fulfilled to full satisfaction
- Pass: at most one bullet point from the above missing

- Plagiarism will have serious consequences

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)

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

- 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