Google DeepMind

Are generative video models the path towards solving visual intelligence?

Robert Geirhos

June 03, 2026

Visual General Intelligence workshop @ CVPR 2026

Based on “Video models are zero-shot learners and reasoners”.
Joint work with my amazing collaborators:


Thaddäus Wiedemer


Effie Li


Paul Vicol

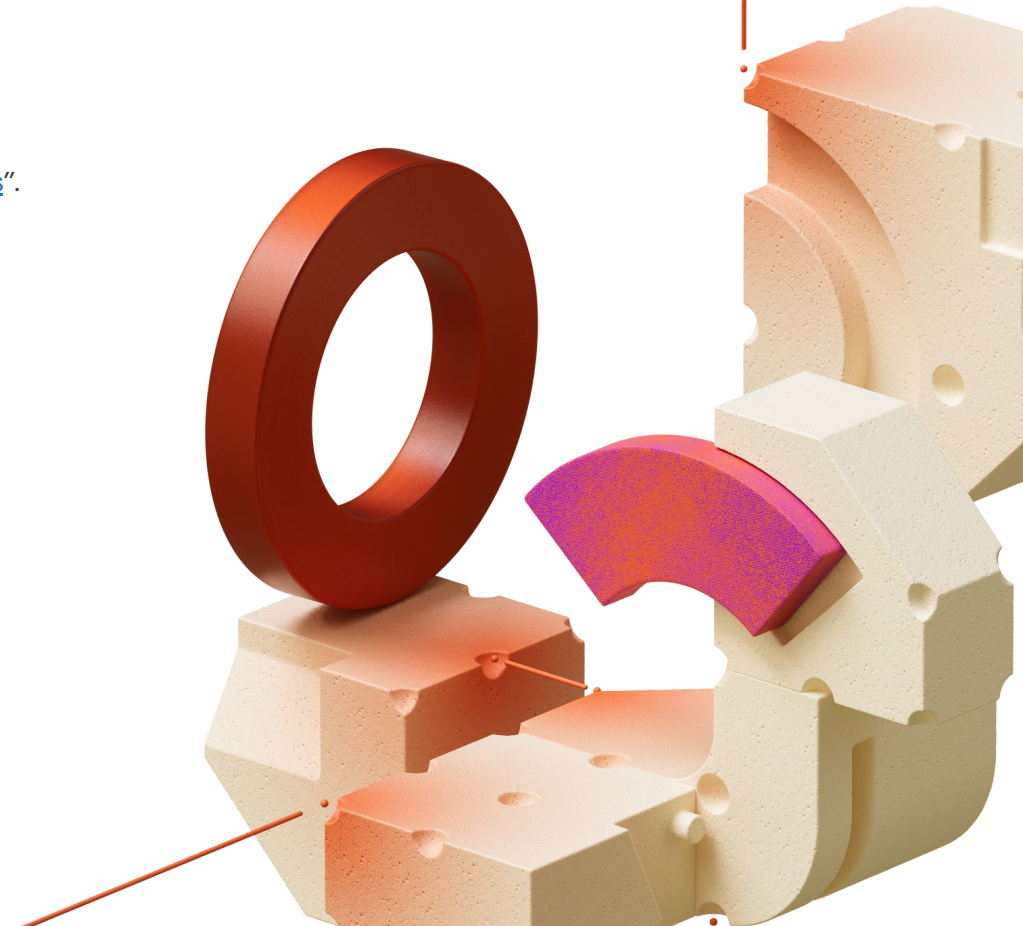

Shane Gu


Nick Matarese

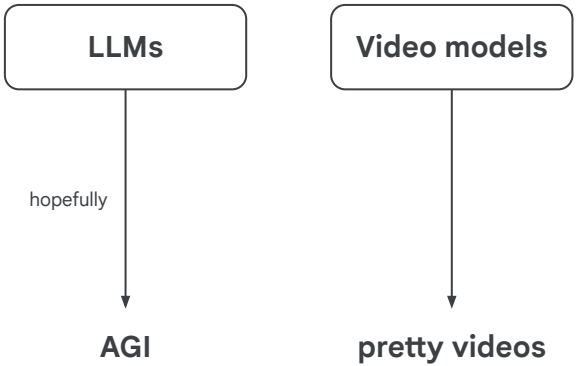

Kevin Swersky


Been Kim

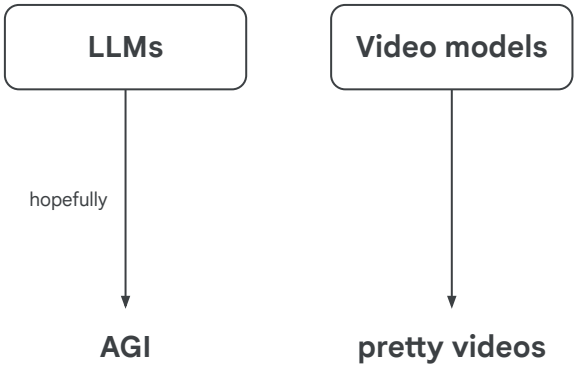

Priyank Jaini



Status quo: division of labor



Status quo: division of labor



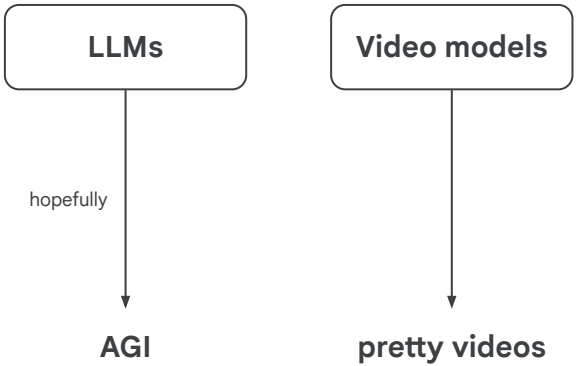
The future?

LLMs
general-purpose
language understanding

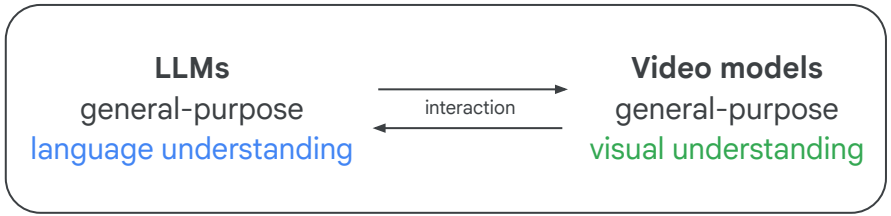
Video models
general-purpose
visual understanding

see also Yang et al. 2024, Video as the New Language for Real-World Decision Making

Status quo: division of labor

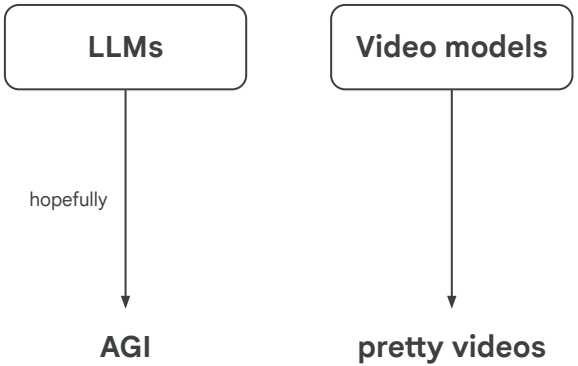


The future?

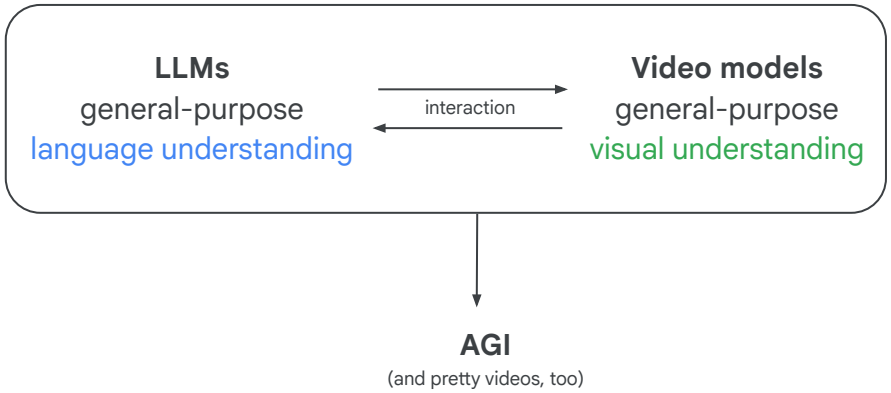


see also Yang et al. 2024, Video as the New Language for Real-World Decision Making

Status quo: division of labor



The future?



see also Yang et al. 2024, Video as the New Language for Real-World Decision Making

The road to general-purpose understanding

Blueprint: the LLM (r)evolution...

Emergent behavior
of larger and larger models

Specialized models
for each and every task

Fine-tuning a base model
for specific tasks

Few-shot in-context learning

Zero-shot learning

The road to general-purpose understanding

Blueprint: the LLM (r)evolution...

Emergent behavior
of larger and larger models

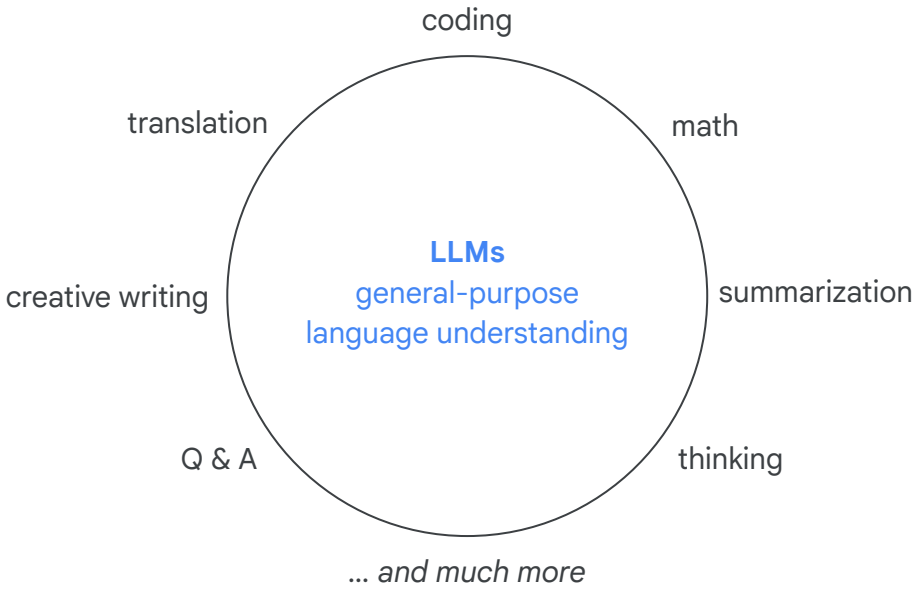
Specialized models
for each and every task

Fine-tuning a base model
for specific tasks

Few-shot in-context learning

Zero-shot learning

...enabled



The road to general-purpose understanding

Vision models aren't there yet...

Specialized models
for each and every task

Fine-tuning a base model
for specific tasks

Few-shot in-context learning

Zero-shot learning

The road to general-purpose understanding

Vision models aren't there yet...

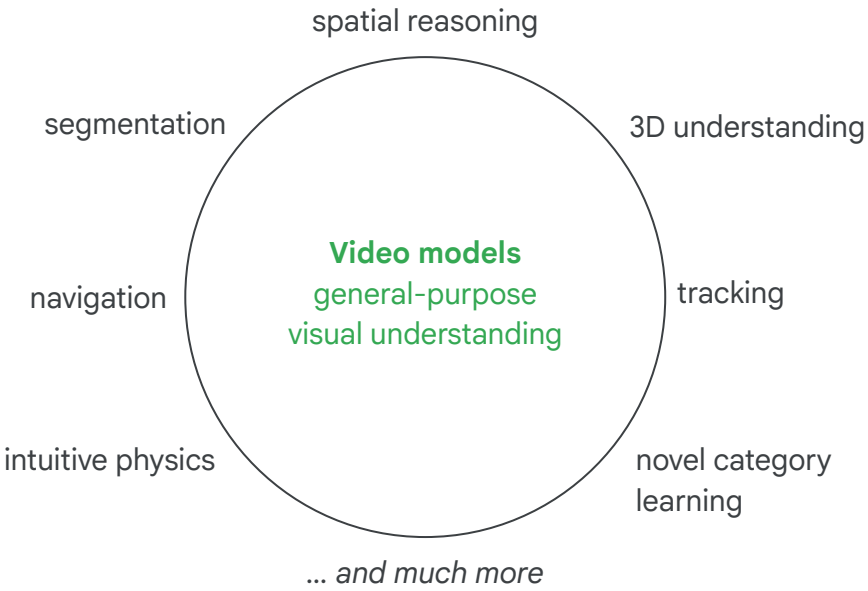
Specialized models
for each and every task

Fine-tuning a base model
for specific tasks

Few-shot in-context learning

Zero-shot learning

...but could enable



“

By building a comprehensive evaluation

Veo 2 & 3

We measure the potential of video models to become vision foundation models by quantifying their emergent zero-shot abilities.

able to perform a task despite not being explicitly trained or adapted for that task

Visual intelligence across the vision stack

PERCEPTION Superresolution

Input frame provided to the model

PERCEPTION Superresolution

Input frame provided to the model

Prompt:
"Perform superresolution on this image. Static camera perspective, no zoom or pan."

PERCEPTION Superresolution

Video generated by Veo 3

Input frame provided to the model

Prompt:
"Perform superresolution on this image. Static camera perspective, no zoom or pan."

PERCEPTION Superresolution

Video generated by Veo 3

Input frame provided to the model

No fine-tuning, no adaptation, nothing: just prompting the “frozen” model!

Prompt:
“Perform superresolution on this image. Static camera perspective, no zoom or pan.”

PERCEPTION Superresolution

first frame

intermediate frame

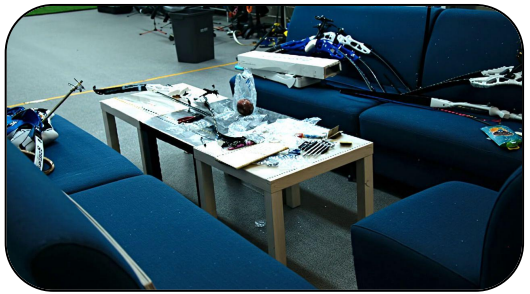
last frame

PERCEPTION Low-light enhancement

“Fully restore the light in this image. Static camera perspective, no zoom or pan.”

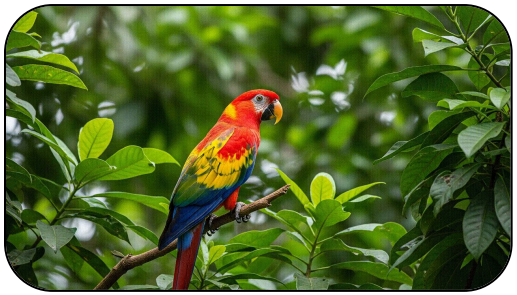
first frame

intermediate frame



last frame

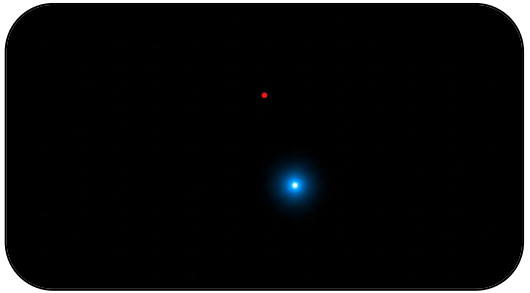
PERCEPTION Keypoint localization



"Add a bright blue dot at the tip of the branch on which the macaw is sitting. The macaw's eye turns bright red. Everything else turns pitch black. Static camera perspective, no zoom or pan"

first frame

intermediate frame



last frame

PERCEPTION Dalmatian illusion understanding

(no text prompt)

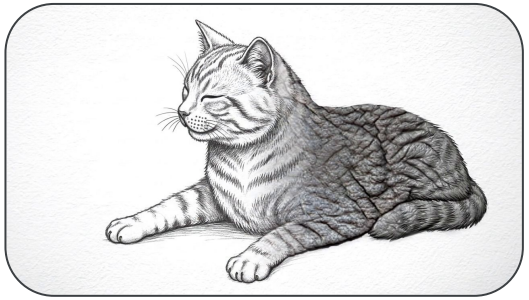
first frame

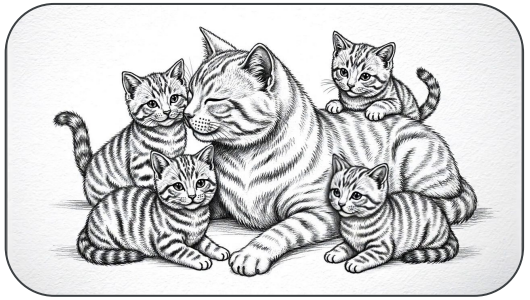
intermediate frame

last frame

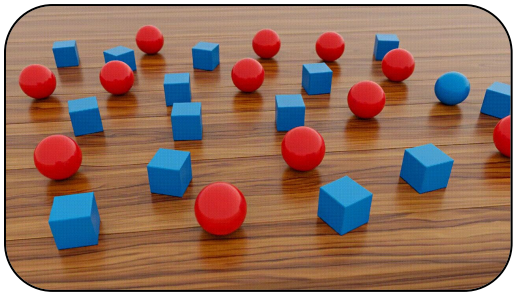
PERCEPTION Shape vs. texture understanding

“Transform the animal in this image into a sketch of the animal surrounded by its family.”

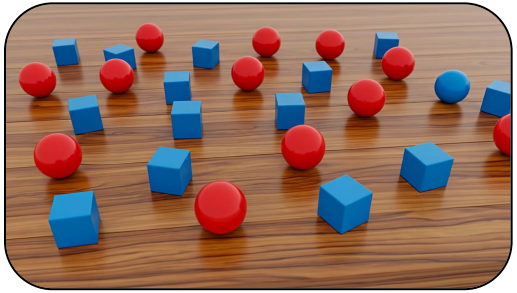


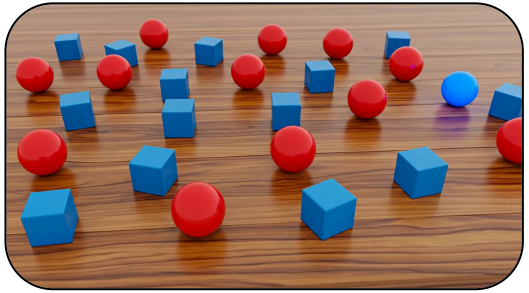


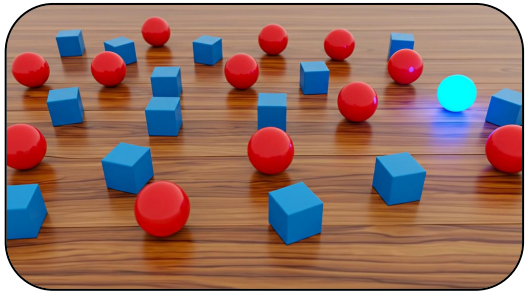
PERCEPTION Conjunctive search



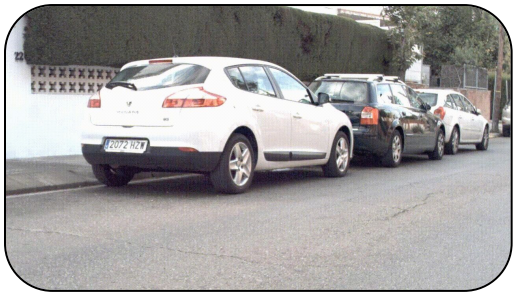
"The blue ball instantly begins to glow. Static camera perspective, no zoom no pan no movement no dolly no rotation."



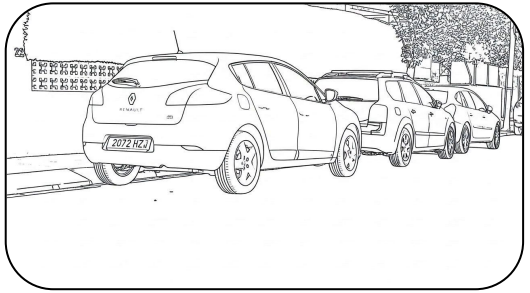




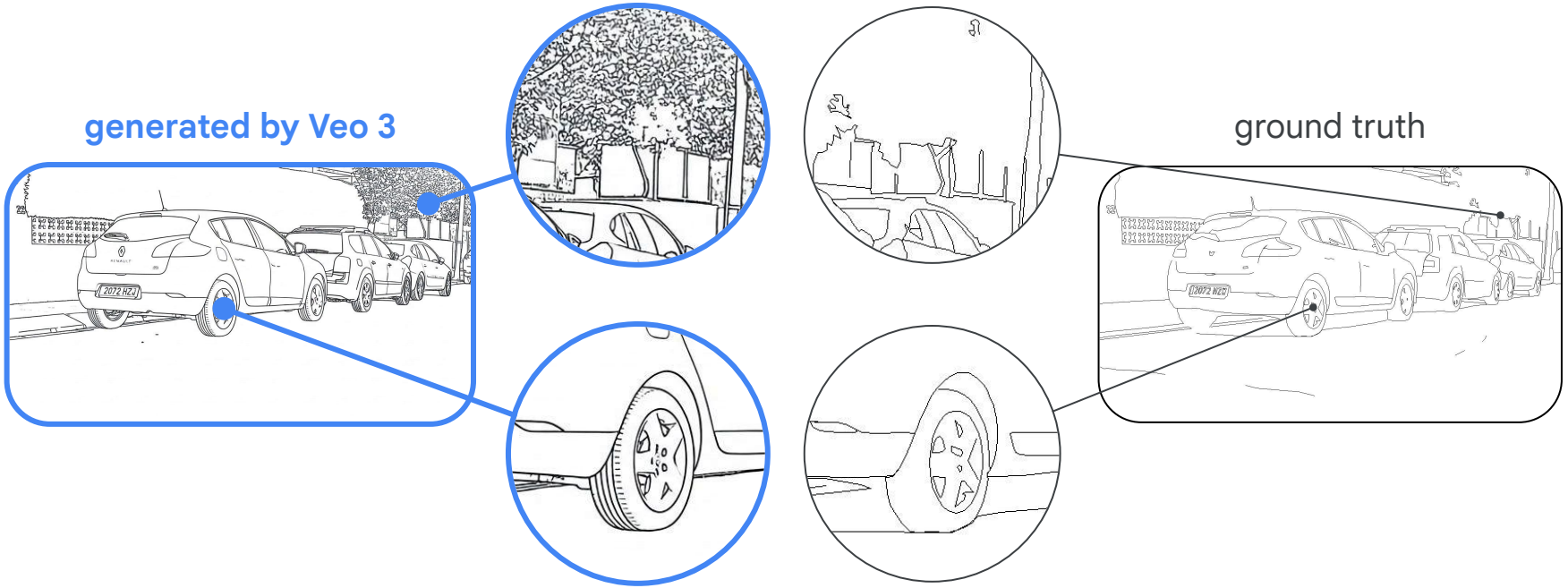
PERCEPTION Edge detection



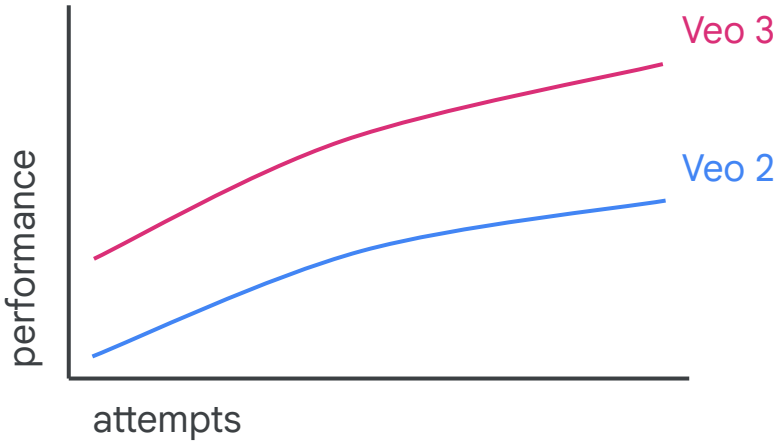
"All edges in this image become more salient by transforming into black outlines. Then, all objects fade away, with just the edges remaining on a white background. Static camera perspective, no zoom or pan."



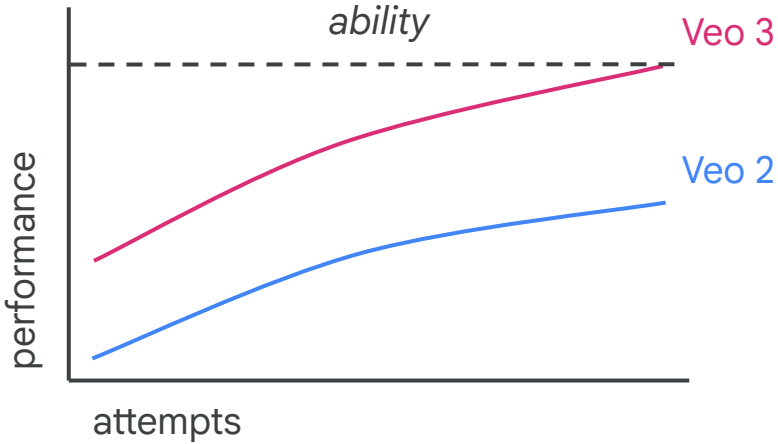
PERCEPTION Edge detection



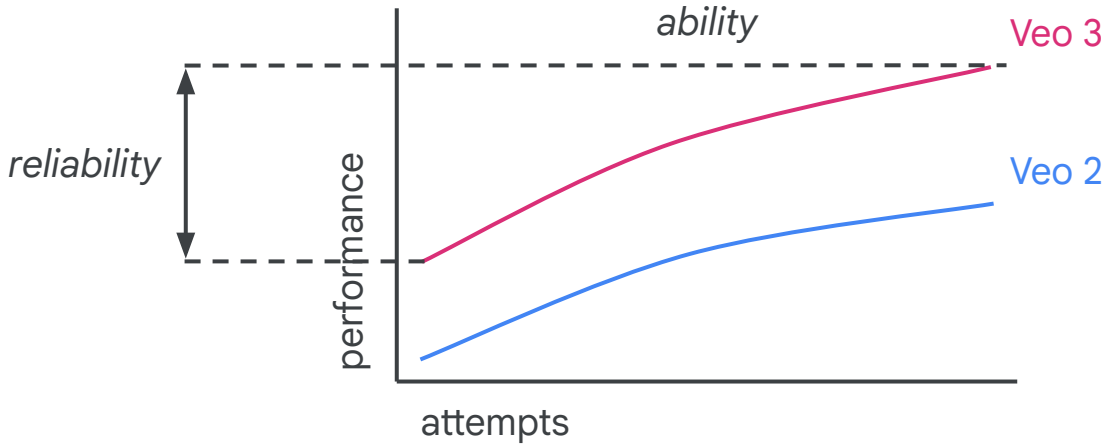
Anatomy of an eval



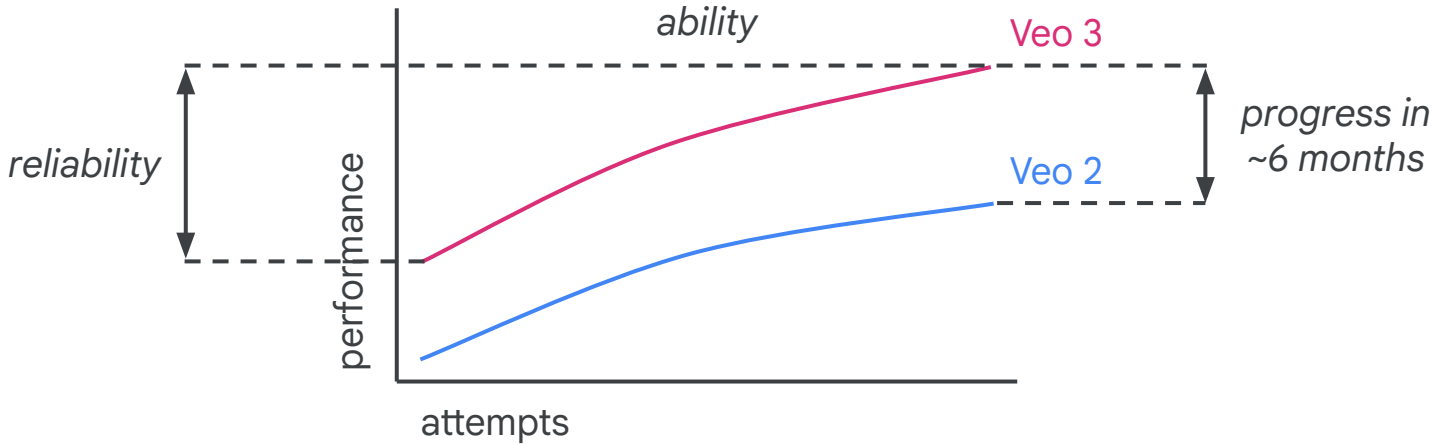
Anatomy of an eval



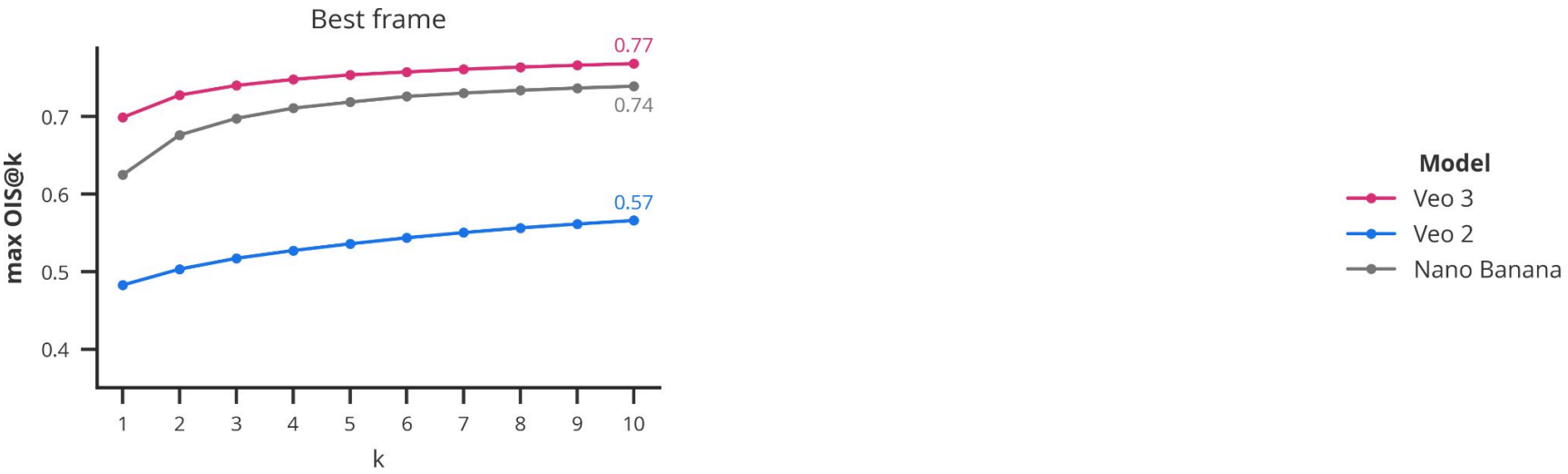
Anatomy of an eval



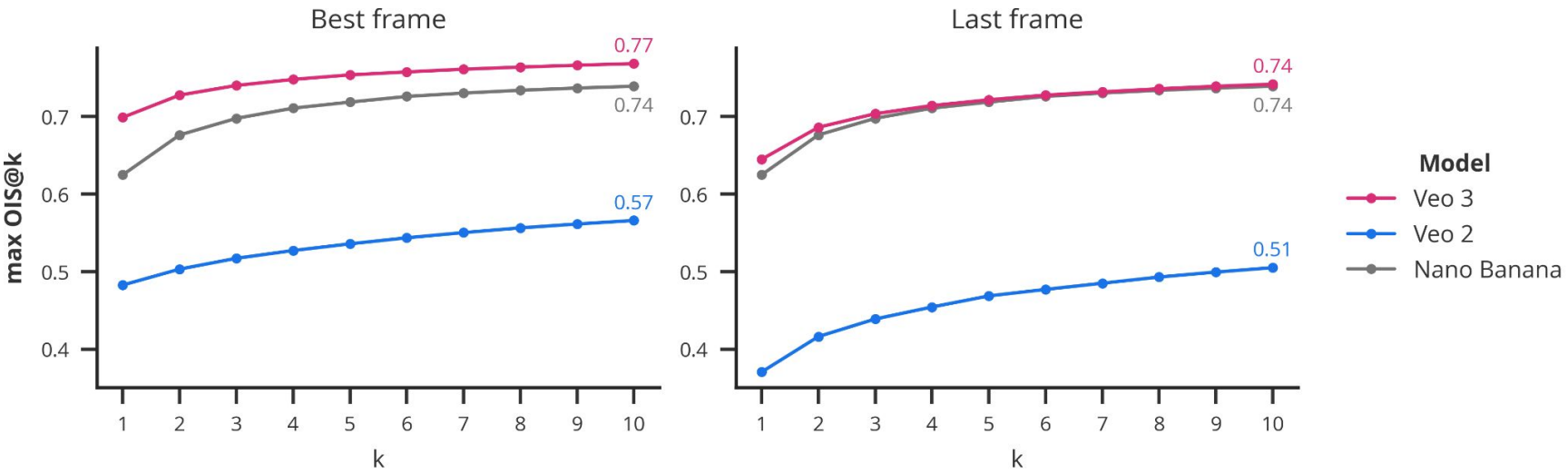
Anatomy of an eval



PERCEPTION Edge detection



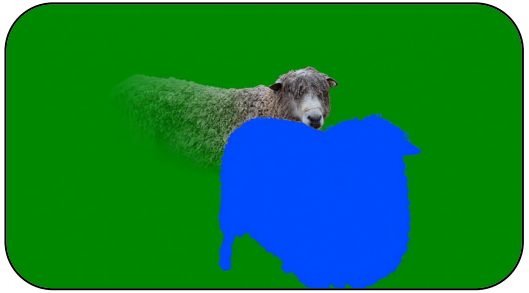
PERCEPTION Edge detection

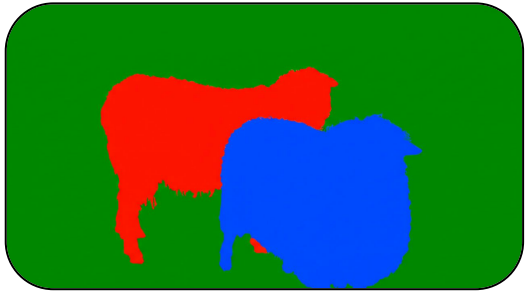


PERCEPTION Segmentation

"Create an animation of instance segmentation being performed on this photograph: each distinct entity is overlaid in a different flat color.
Scene:

- *The animation starts from the provided, unaltered photograph.*
- *The scene in the photograph is static and doesn't move.*
- *First, the background fades to green.*
- *Then, the first entity is covered by a flat color ...*

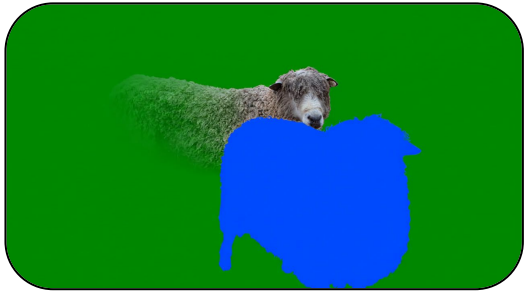


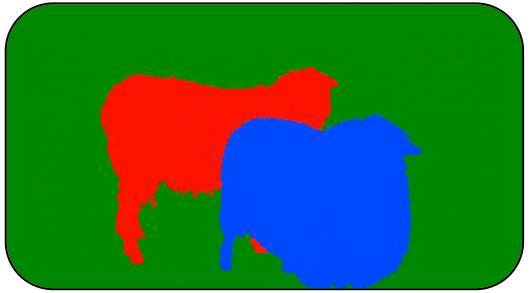


PERCEPTION Segmentation

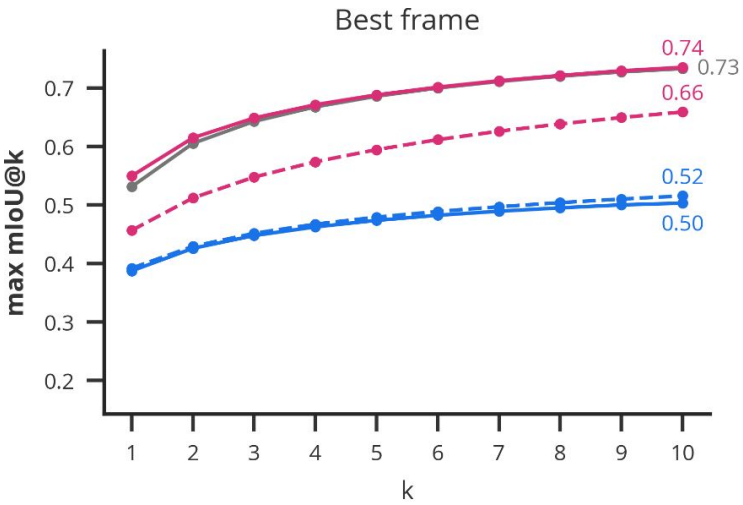
"...
*First, the background fades to **green**.*
..."

visual prompt supports **verbal prompt**





PERCEPTION Segmentation



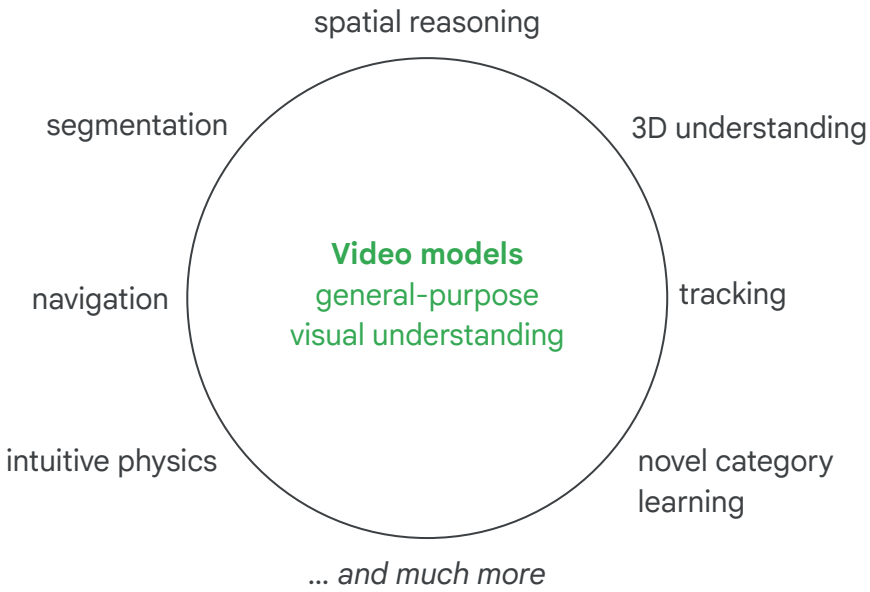
Model

- Veo 3
- Veo 2
- Nano Banana

Background

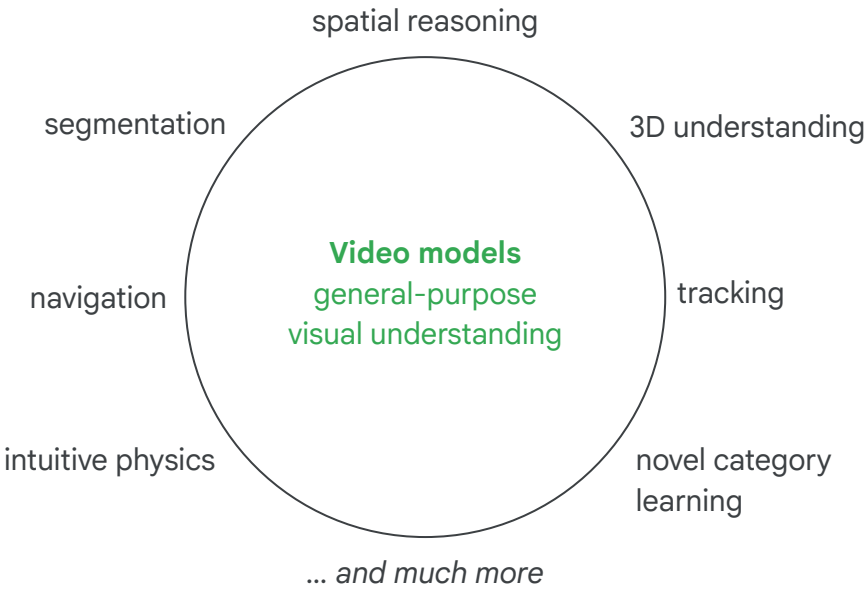
- green
- white

PERCEPTION: QUO VADIS, COMPUTER VISION?



PERCEPTION: QUO VADIS, COMPUTER VISION?

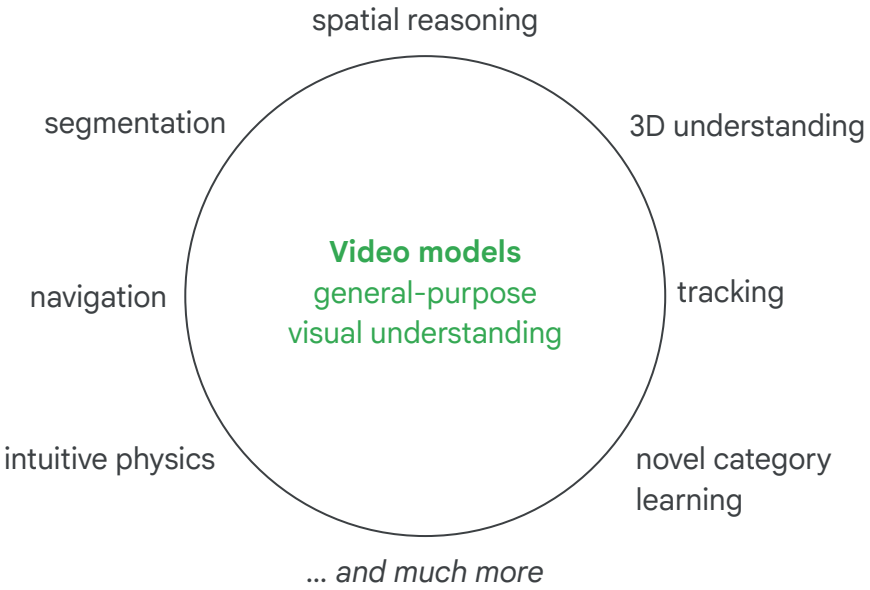
Just like LLMs replaced task-specific NLP models, **video models will likely replace most bespoke models in computer vision**—once they become sufficiently cheap and reliable.



PERCEPTION: QUO VADIS, COMPUTER VISION?

Just like LLMs replaced task-specific NLP models, **video models will likely replace most bespoke models in computer vision**—once they become sufficiently cheap and reliable.

Instead of task-specific training, we will just prompt models to solve visual tasks.



Visual intelligence across the vision stack

MODELING

Based on their *perception* of the world, video models are starting to *model* the visual world, too.

MODELING

Based on their *perception* of the world, video models are starting to *model* the visual world, too.

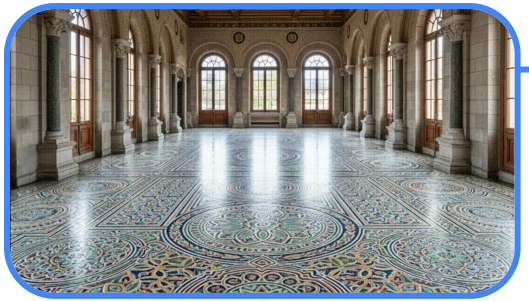
Emergent area with lots of progress in terms of benchmarks & evals:

- **IntPhys** Riochet, Castro, Bernard, Lerer, Fergus, Izard & Dupoux (2018)
- **Physion++** Tung, Ding, Chen, Bear, Gan, Tenenbaum, Yamins, Fan & Smith 2023)
- **VideoPhy** Bansal, Lin, Xie, Zong, Yarom, Bitton, Jiang, Sun, Chang & Grover (2024)
- **PhysVidBench** Tezcan*, Sanli*, Erdem & Erdem (2025)
- **Physics-IQ** Motamed, Culp, Swersky, Jaini & Geirhos (2025)
- **Morpheus** Zhang, Cherniavskii, Zadaianchuk, Tragoudaras, Vozikis, Nijdam, Prinzhorn, Bodracska, Sebe & Gavves (2025)

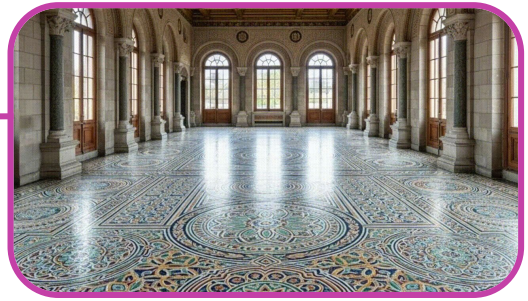
& many more!

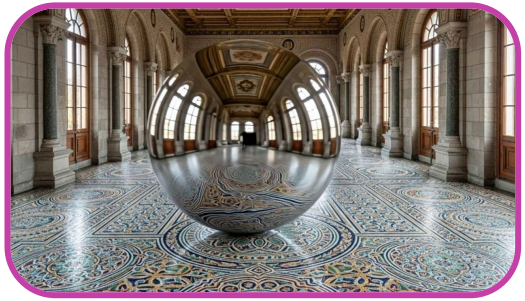
Since this area is already covered well by existing benchmarks, we just investigate qualitative samples here.

MODELING Material optics



*"A giant **glass sphere** / **mirror-polish metal sphere** rolls through the room. Static camera, no pan, no zoom, no dolly"*





MODELING Buoyancy

"The hand lets go of the object. Static camera, no pan, no zoom, no dolly."

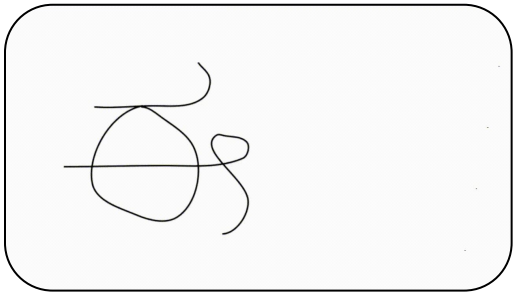
rock

bottle cap

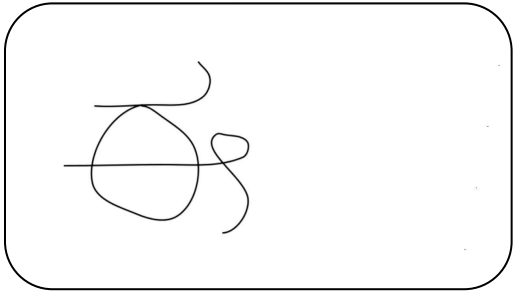
MODELING Memory

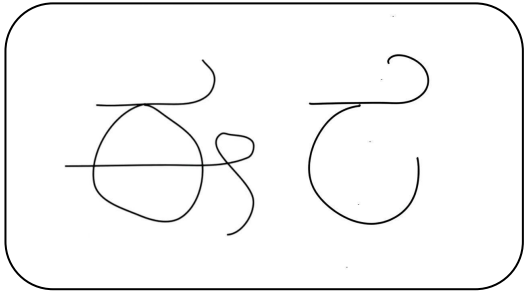
“The camera zooms in to give a close up of the person looking out the window, then zooms back out to return to the original view.”

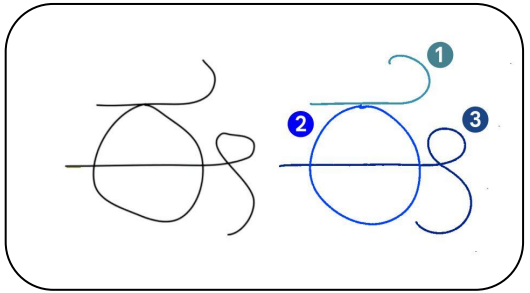
MODELING Parsing into parts



"Stroke-by-stroke, a replica of the symbol is drawn on the right."



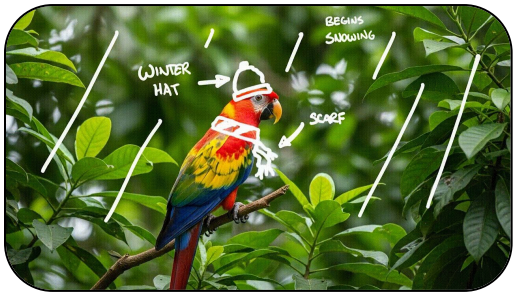




(annotations by us)

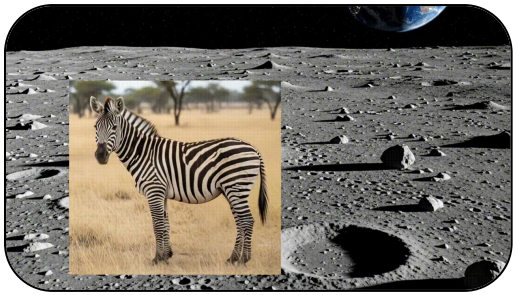
Visual intelligence across the vision stack

MANIPULATION Image editing

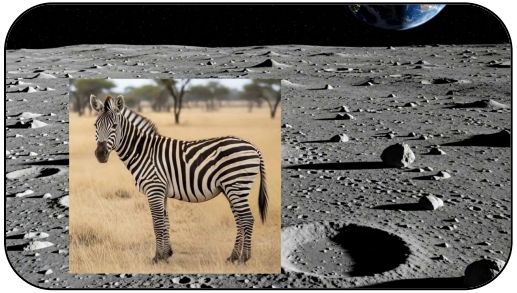


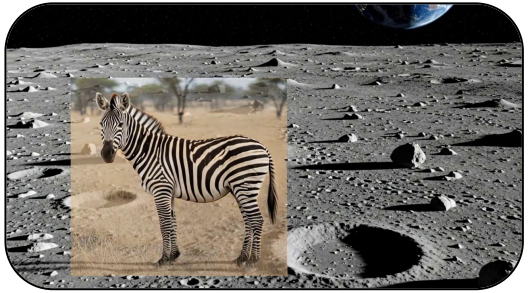
"Changes happen instantly."

MANIPULATION Image editing

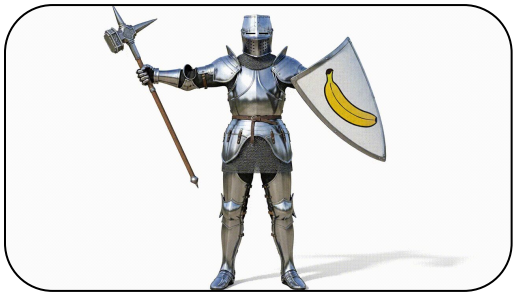


“A smooth animation blends the zebra naturally into the scene, removing the background of the zebra image, so that the angle, lighting, and shading look realistic. The final scene perfectly incorporates the zebra into the scene.”





MANIPULATION Image editing (reposing)

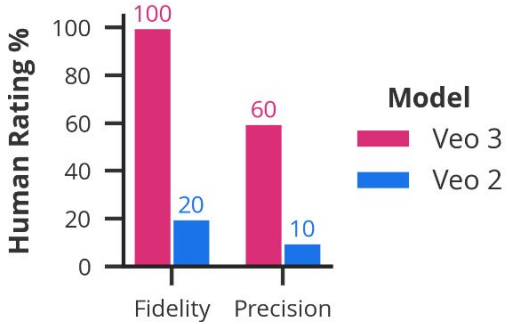


"The knight turns to face to the right and drops on one knee, lifting the shield above his head to protect himself and resting the hilt of his weapon on the ground."

MANIPULATION Image editing

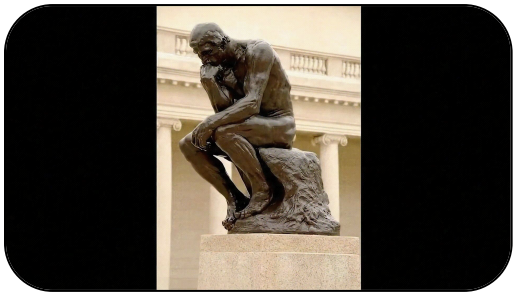
Task-specific prompt

*"Create a smooth, static animation that slowly **turns the fire hydrant red**. Do not change anything else. No zoom, no pan, no dolly."*

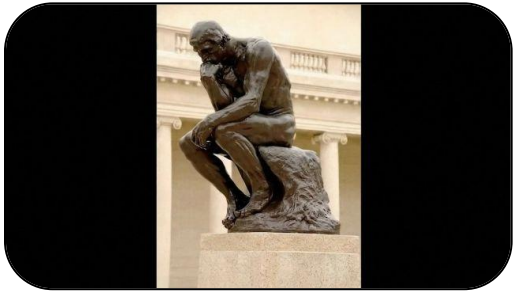


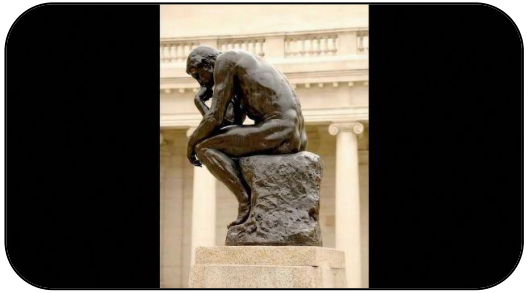
Evaluated on samples from the Emu-Edit dataset (Sheynin et al. 2023)

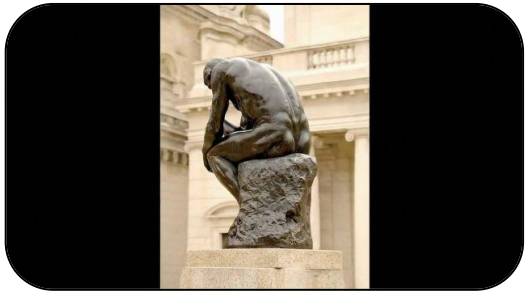
MANIPULATION Novel-view synthesis



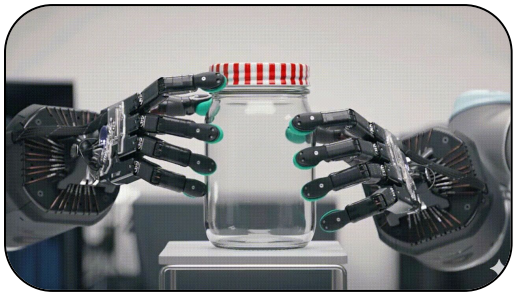
“Create a smooth, realistic animation where the camera seems to rotate around the object showing the object from all the sides. Do not change anything else. No zoom. No pan.”



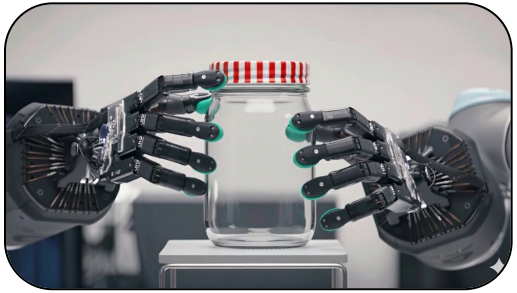


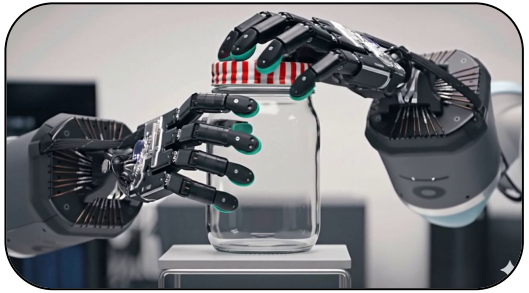


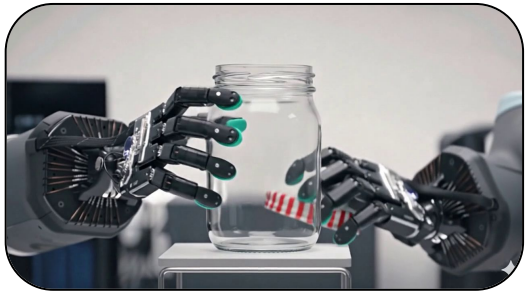
MANIPULATION Dexterous manipulation



“Use common sense and have the two robot hands attached to robot arms open the jar, like how a human would.”

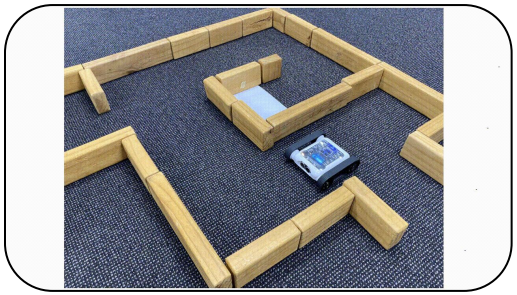




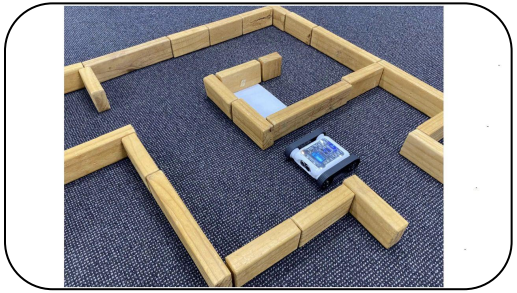


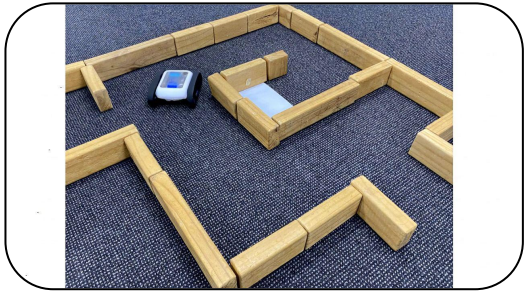
Visual intelligence across the vision stack

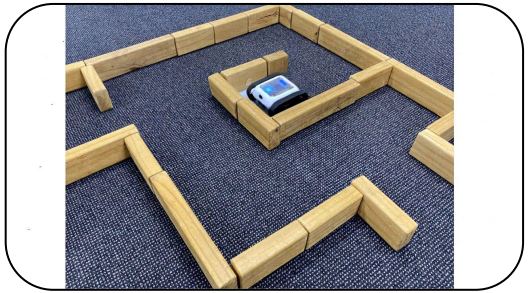
REASONING Robot navigation



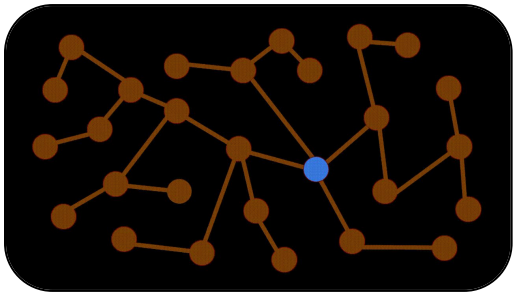
“The robot drives to the blue area. Static camera perspective, no movement no zoom no scan no pan.”



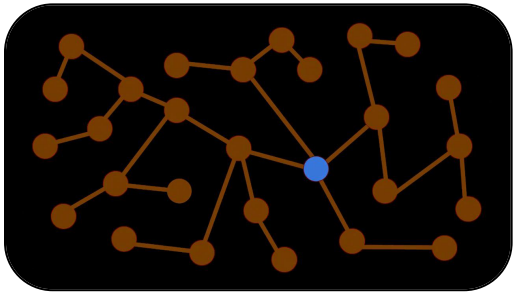


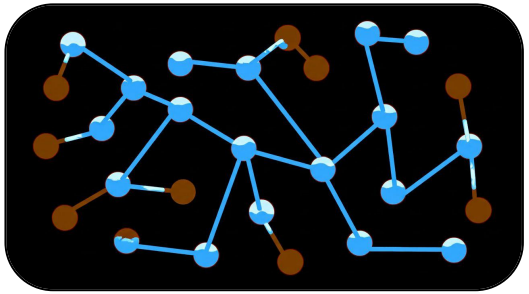


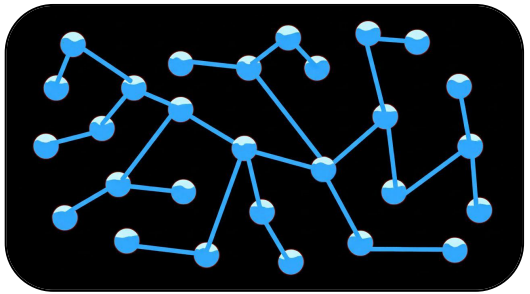
REASONING Graph traversal



"Starting from the blue well, an unlimited supply of blue water moves through the connected channel system without spilling into the black area."



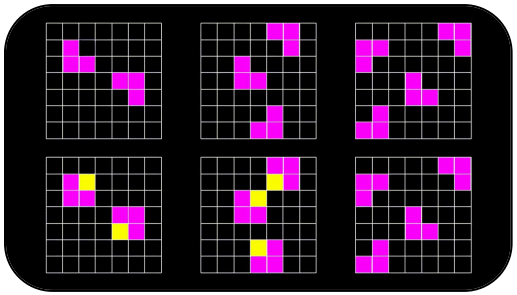




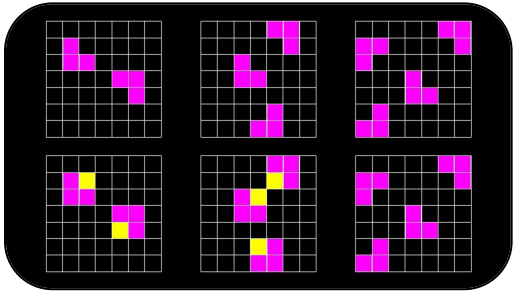
REASONING Tool use

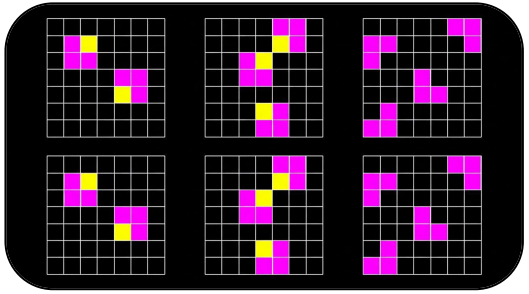
"A person retrieves the walnut from the aquarium."

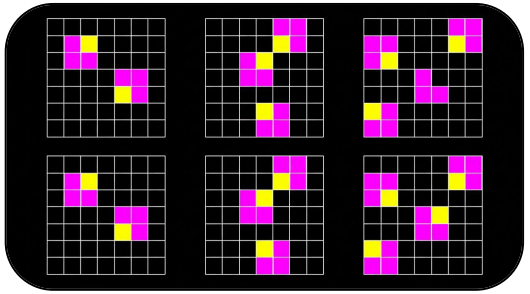
REASONING Rule extrapolation



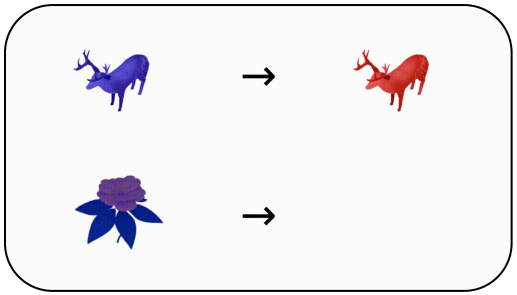
“Modify the lower-right grid to adhere to the rule established by the other grids. You can fill cells, clear cells, or change a cell’s color. Only modify the lower-right grid, don’t modify any of the other grids. Static scene, no zoom, no pan, no dolly.”



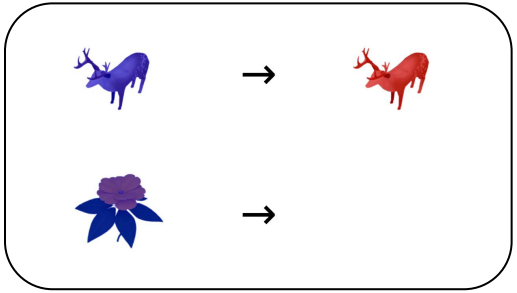


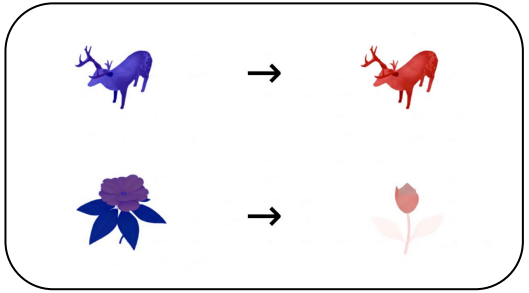


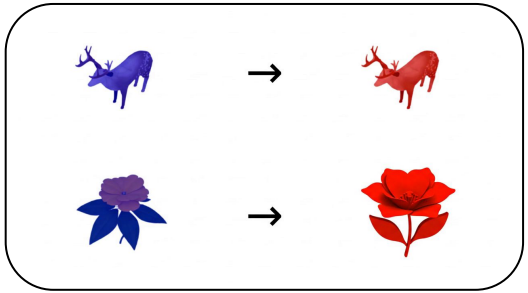
REASONING Visual analogies



"Create a smooth animation to generate the missing object in the lower right region and solve the visual analogy. The original three objects must remain still. Static shot, no zoom no pan no dolly."

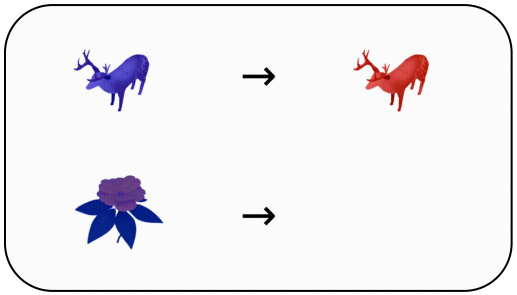




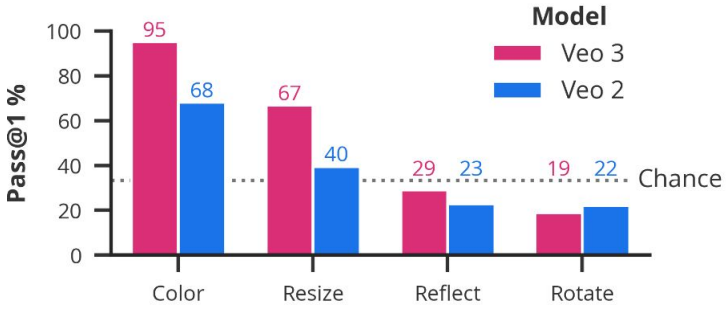


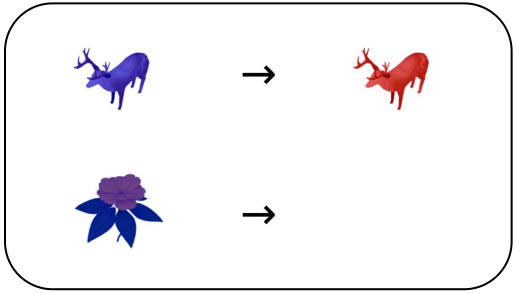
Evaluated on samples from the KIVA dataset (Yiu et al. 2024)

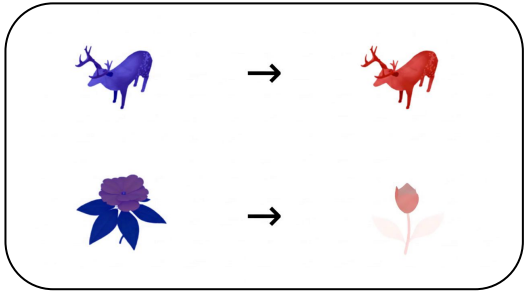
REASONING Visual analogies

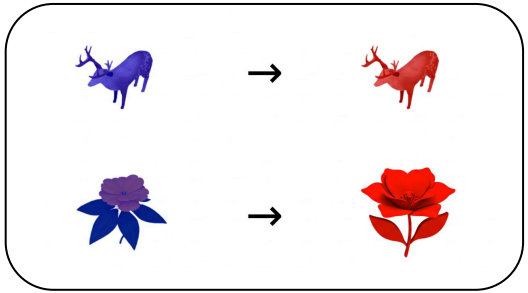


"Create a smooth animation to generate the missing object in the lower right region and solve the visual analogy. The original three objects must remain still. Static shot, no zoom no pan no dolly."



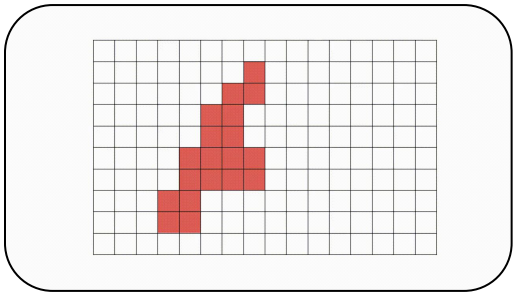




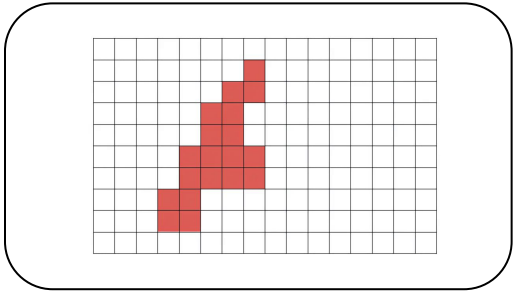


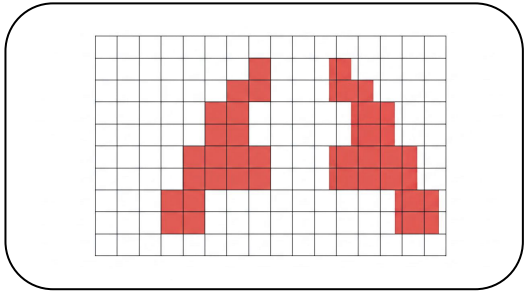
Evaluated on samples from the KIVA dataset (Yiu et al. 2024)

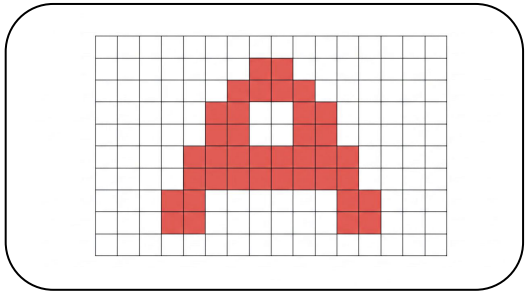
REASONING Visual symmetry



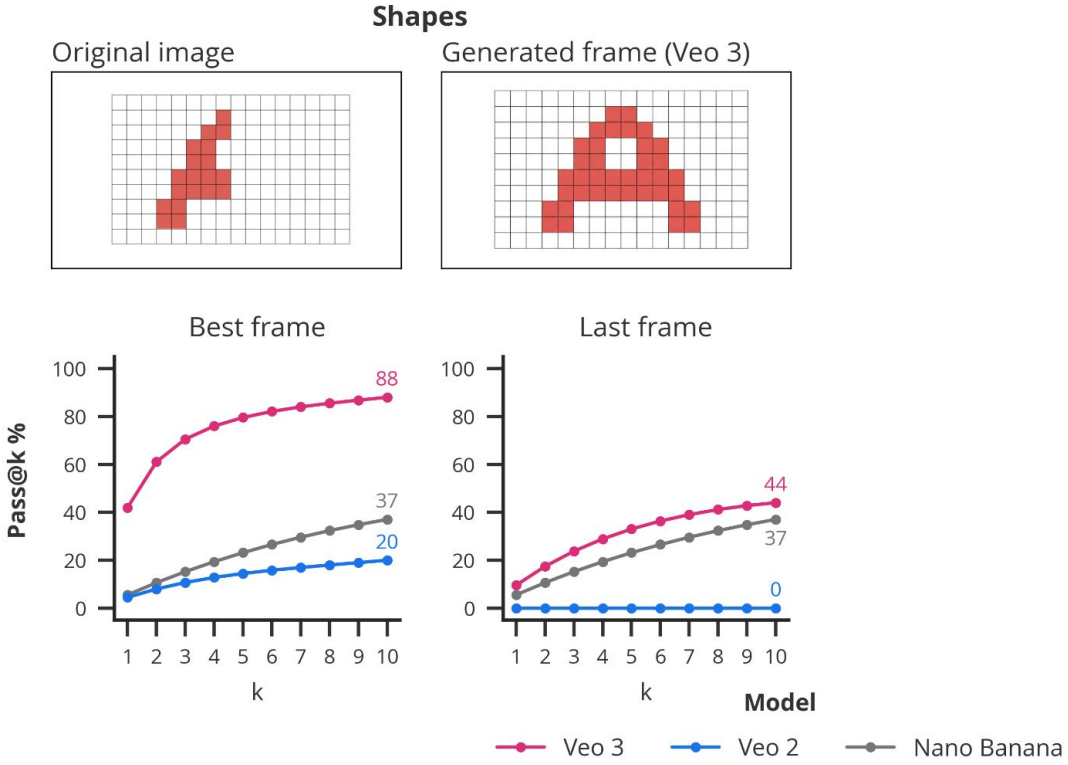
"Instantly reflect this pattern along the central, vertical axis while keeping the existing colored pattern without modification. Static camera perspective, no zoom or pan."



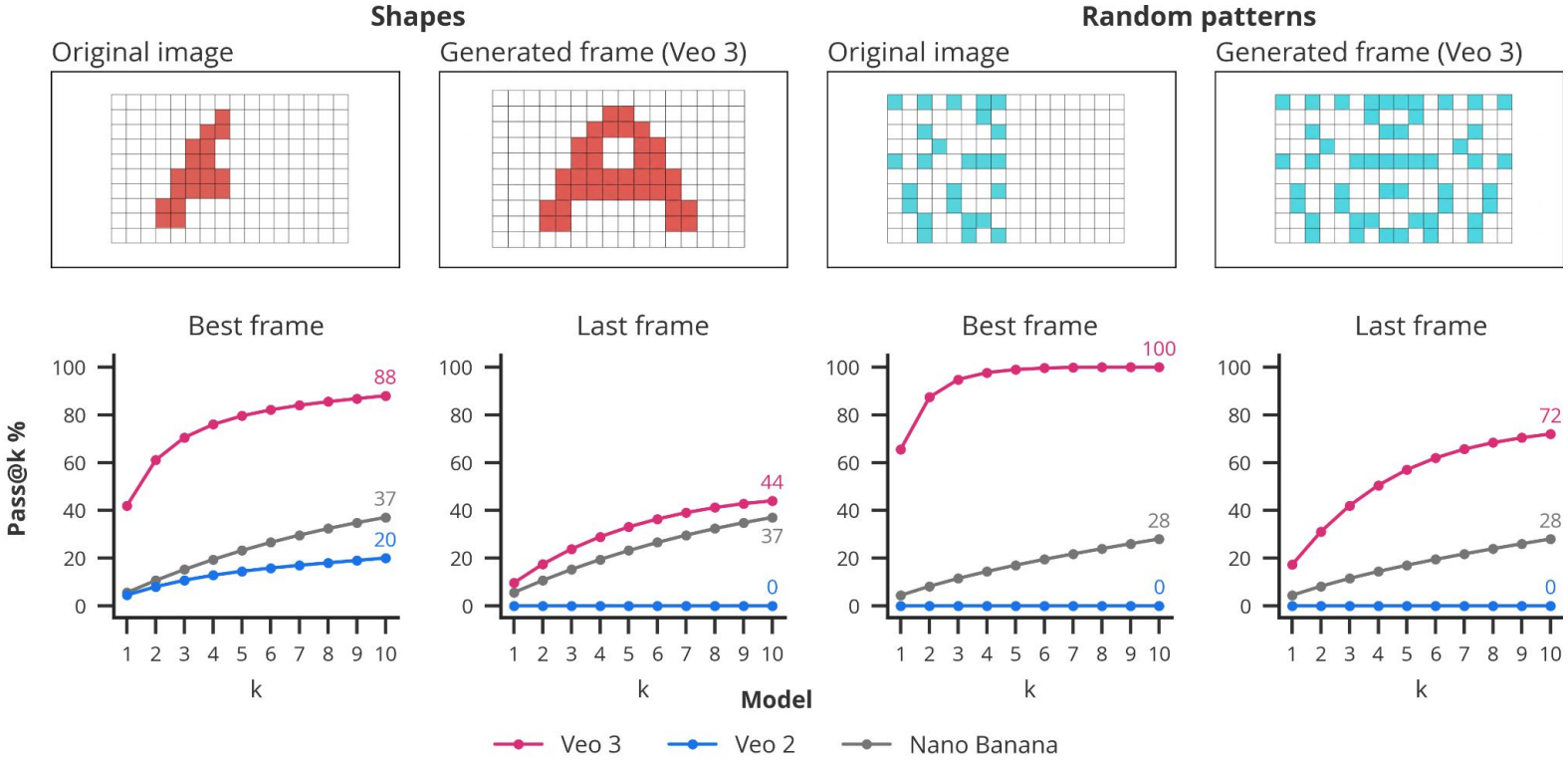




REASONING Visual symmetry

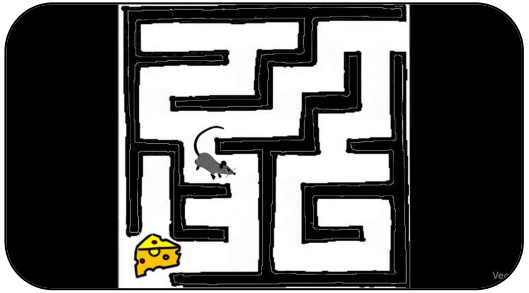


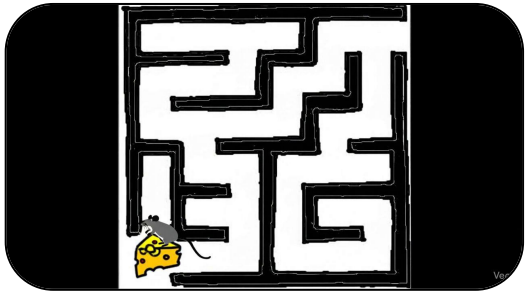
REASONING Visual symmetry



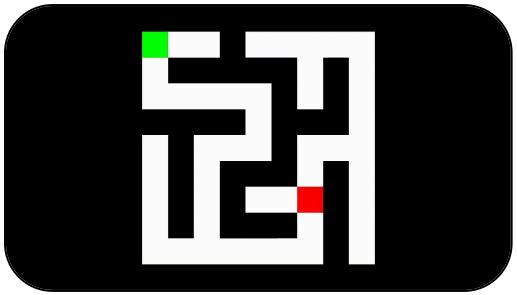
REASONING Maze solving

“Without crossing any black boundary, the grey mouse from the corner skillfully navigates the maze by walking around until it finds the yellow cheese.”



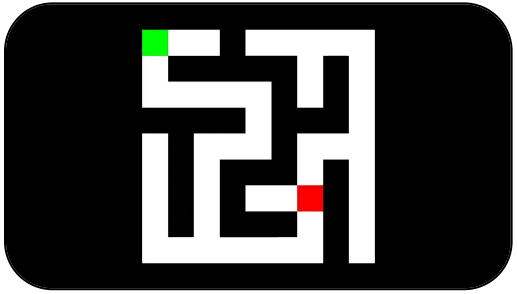


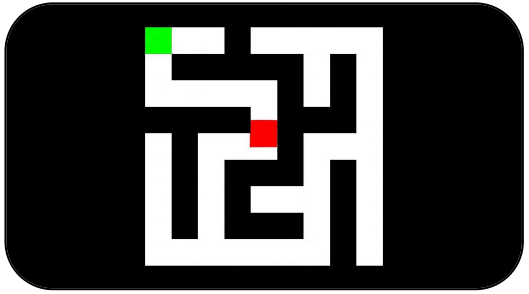
REASONING Maze solving

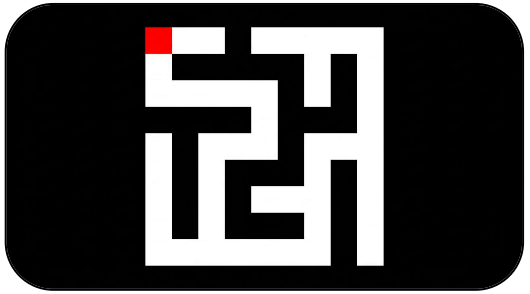


"Create a 2D animation based on the provided image of a maze. The red square slides smoothly along the white path, stopping perfectly on the green square. The red square never slides or crosses into the black areas of the maze. The camera is a static, top-down view showing the entire maze.
Maze:

- *The maze paths are white, the walls are black.*
- *The red square moves to the goal position, represented by a green square...*

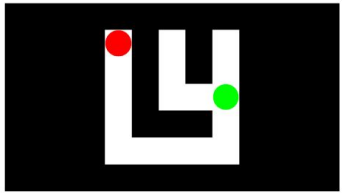


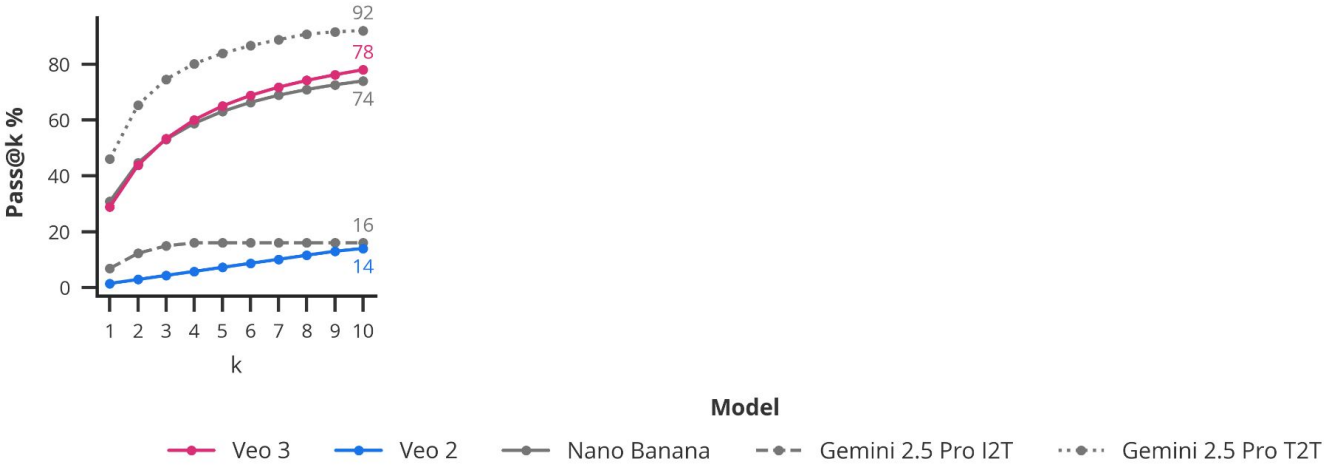




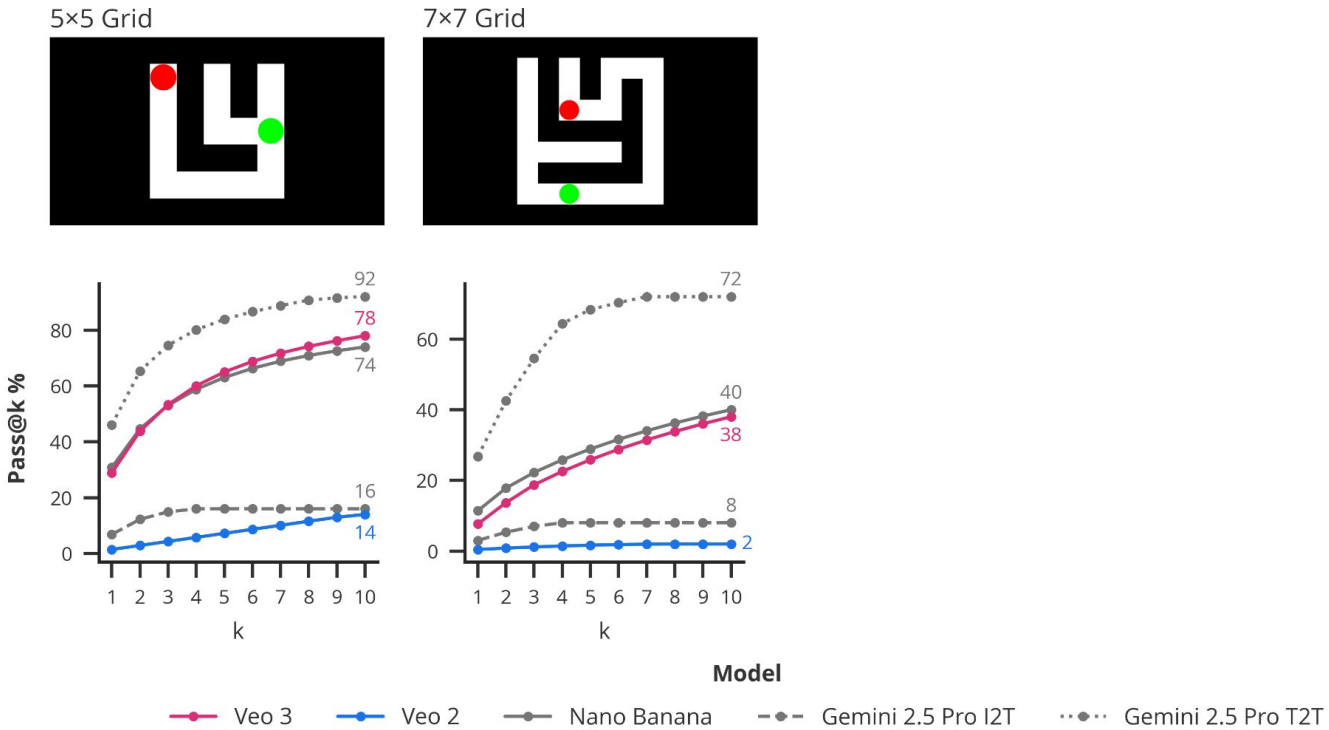
REASONING Maze solving

5×5 Grid

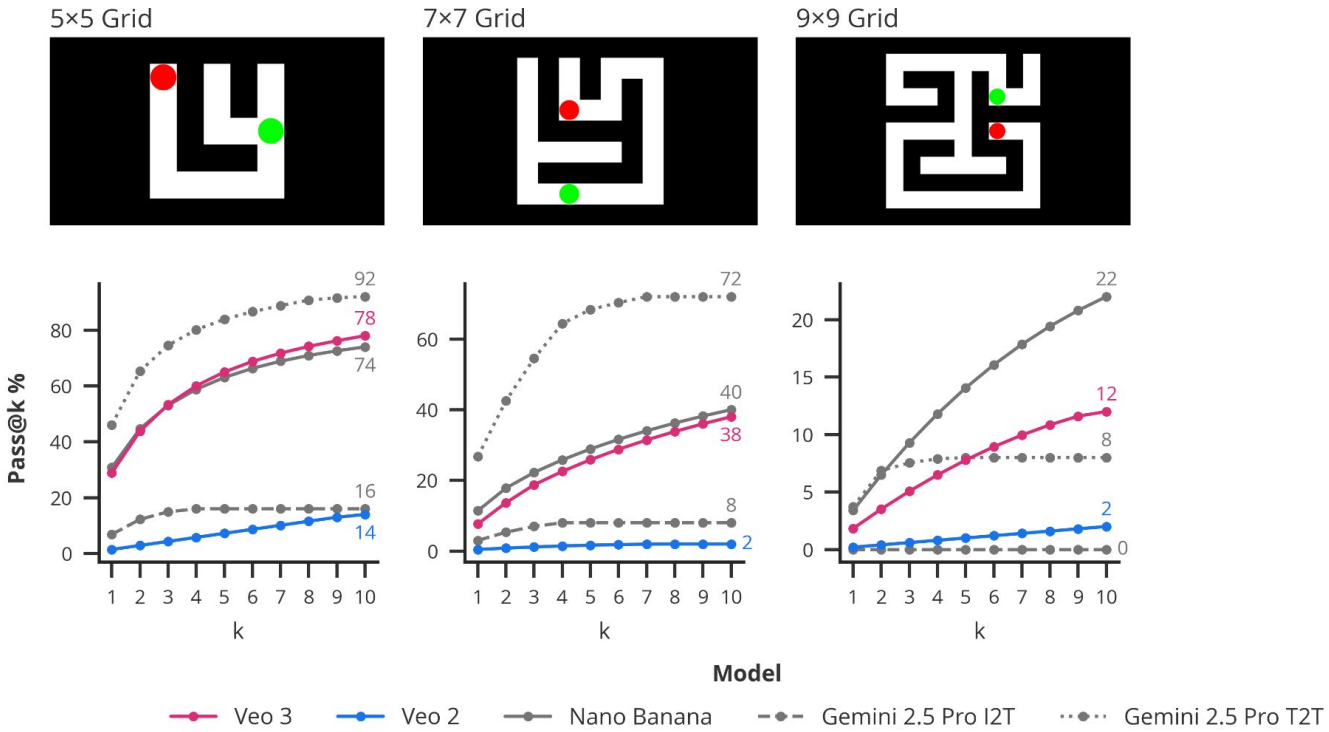




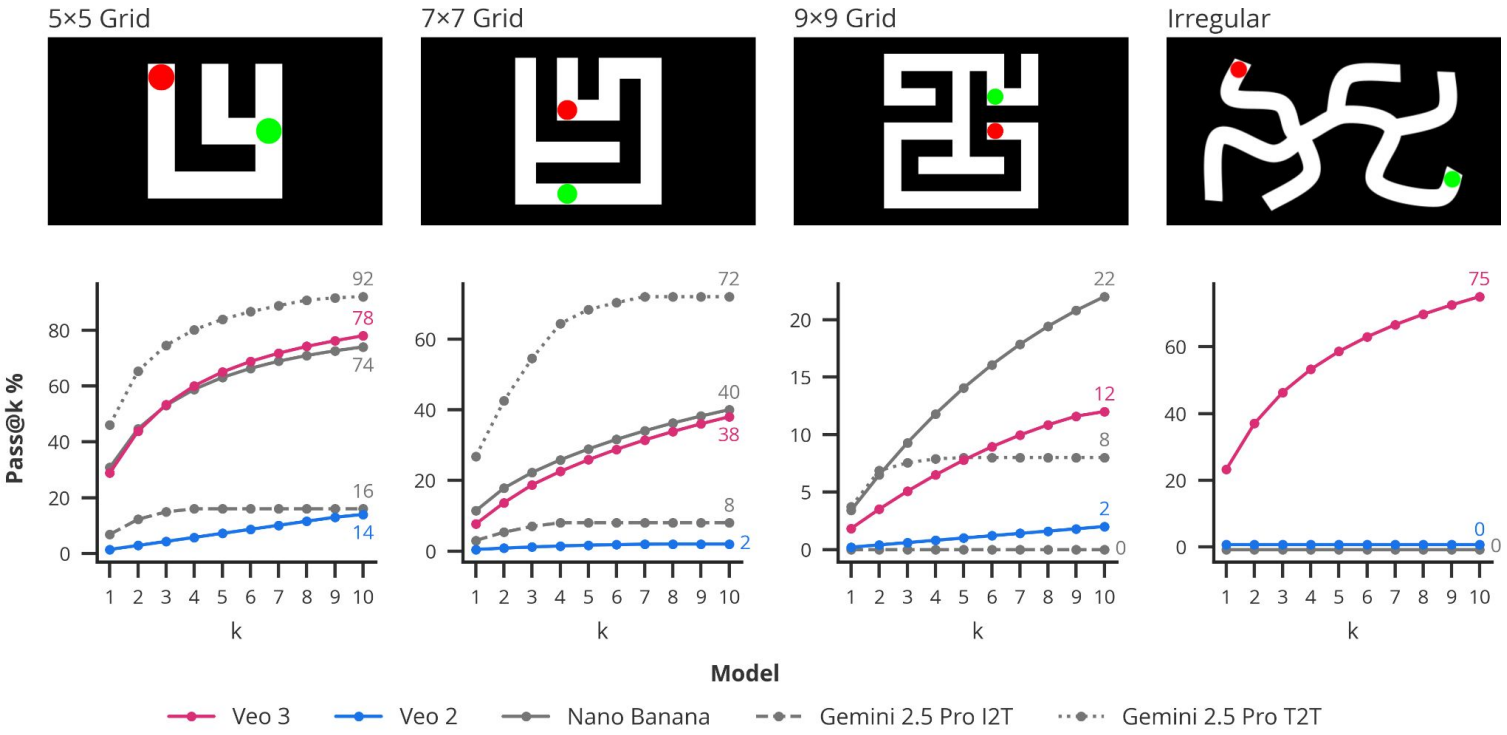
REASONING Maze solving



REASONING Maze solving



REASONING Maze solving



REASONING across diffusion steps

~ “move plant to the left”

Not our work, this is from a recent paper:
Wang, Cai, Pu, Xu, Yin, Wang, Ji, Gu, Li, Huang, Deng, Lin, Liu, & Yang. "Demystifing Video Reasoning", arXiv preprint arXiv:2603.16870 (2026).

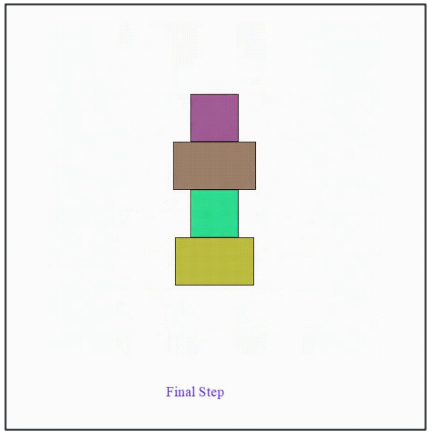
REASONING across diffusion steps

~ “move plant to the left”

Not our work, this is from a recent paper:
Wang, Cai, Pu, Xu, Yin, Wang, Ji, Gu, Li, Huang, Deng, Lin, Liu, & Yang. "Demystifing Video Reasoning", arXiv preprint arXiv:2603.16870 (2026).

REASONING across diffusion steps

~ “move plant to the left”

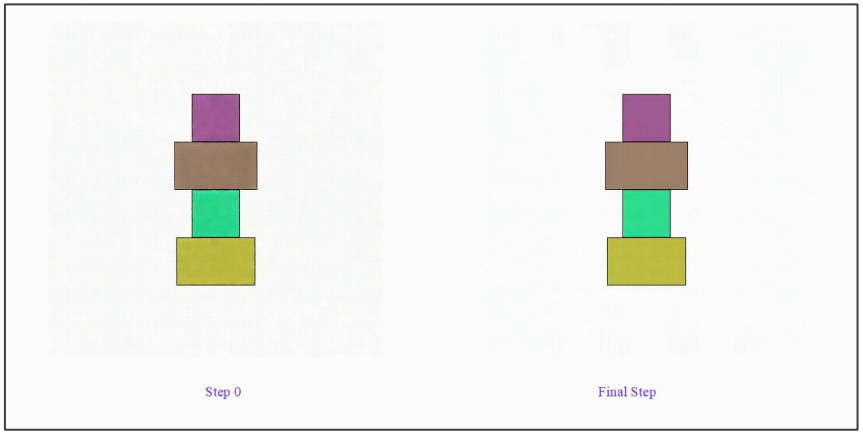


“remove the shapes one by one from top to bottom”

Not our work, this is from a recent paper:
Wang, Cai, Pu, Xu, Yin, Wang, Ji, Gu, Li, Huang, Deng, Lin, Liu, & Yang. "Demystifing Video Reasoning", arXiv preprint arXiv:2603.16870 (2026).

REASONING across diffusion steps

~ “move plant to the left”



“remove the shapes one by one from top to bottom”

Not our work, this is from a recent paper:
Wang, Cai, Pu, Xu, Yin, Wang, Ji, Gu, Li, Huang, Deng, Lin, Liu, & Yang. "Demystifing Video Reasoning", arXiv preprint arXiv:2603.16870 (2026).

Video models are **chain-of-frames / -steps** reasoners across space and time.

This parallels chain-of-thought reasoning in language models.

Video models are **chain-of-frames / -steps** reasoners across space and time.

This parallels chain-of-thought reasoning in language models.

Chain of thought (LLMs):

input → intermediate steps (symbolic reasoning) → output

Video models are **chain-of-frames / -steps** reasoners across space and time.

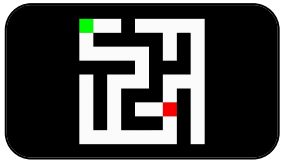
This parallels chain-of-thought reasoning in language models.

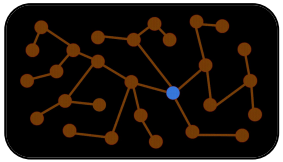
Chain of thought (LLMs):

input → intermediate steps (symbolic reasoning) → output

Chain of frames (video models):

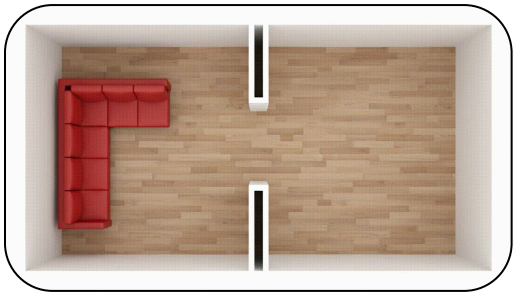
input → intermediate steps (visual reasoning) → output



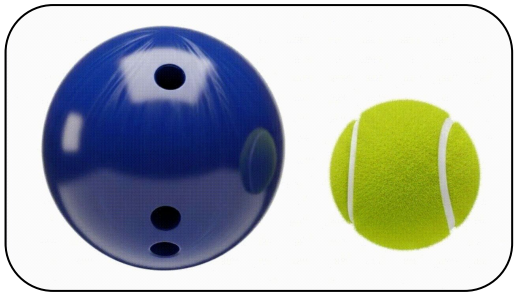


Failure cases

FAIL Implausible object interactions



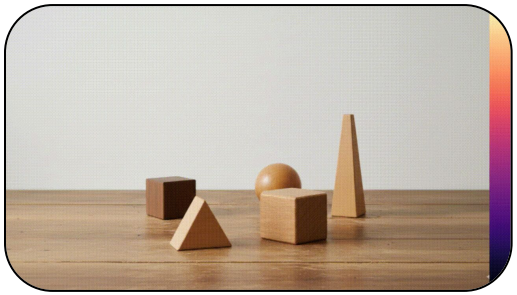
“The red couch slides from the left room over into the right room, skillfully maneuvering to fit through the doorways without bumping into the walls. The walls are fixed: they don’t shift or disappear, and no new walls are introduced. Static camera, no pan, no zoom, no dolly.”



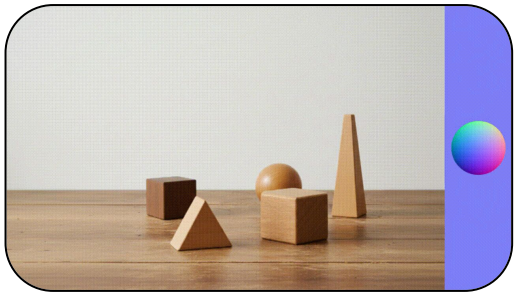
“The two objects collide in slow motion. Static camera, no pan, no zoom, no dolly.”

“A person tries to put the golf ball in the vase. Static camera, no pan, no zoom, no dolly.”

FAIL Complex tasks



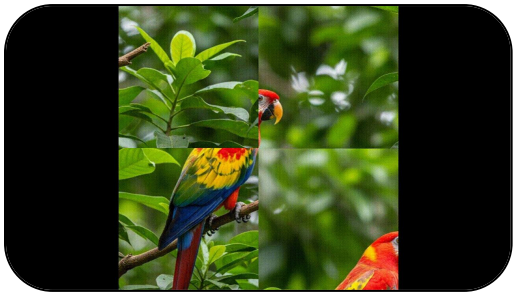
“The image transitions to a depth-map of the scene: Darker colors represent pixels further from the camera, lighter colors represent pixels closer to the camera. The exact color map to use is provided on the right side of the image. Static scene, no pan, no zoom, no dolly.”



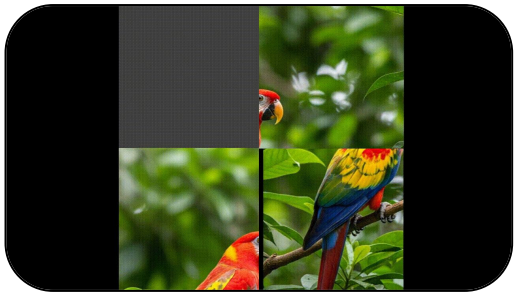
“The image transitions to a surface-normal map of the scene: the red/green/blue color channel specify the direction of the surface-normal at each point, as illustrated on the right side of the image on a sphere. Static scene, no pan, no zoom, no dolly.”

“Generate a video of two metal robotic arms properly folding the t-shirt on the table.”

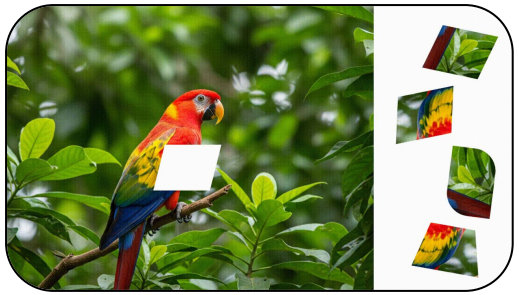
FAIL Puzzles



“Unscramble this image,”

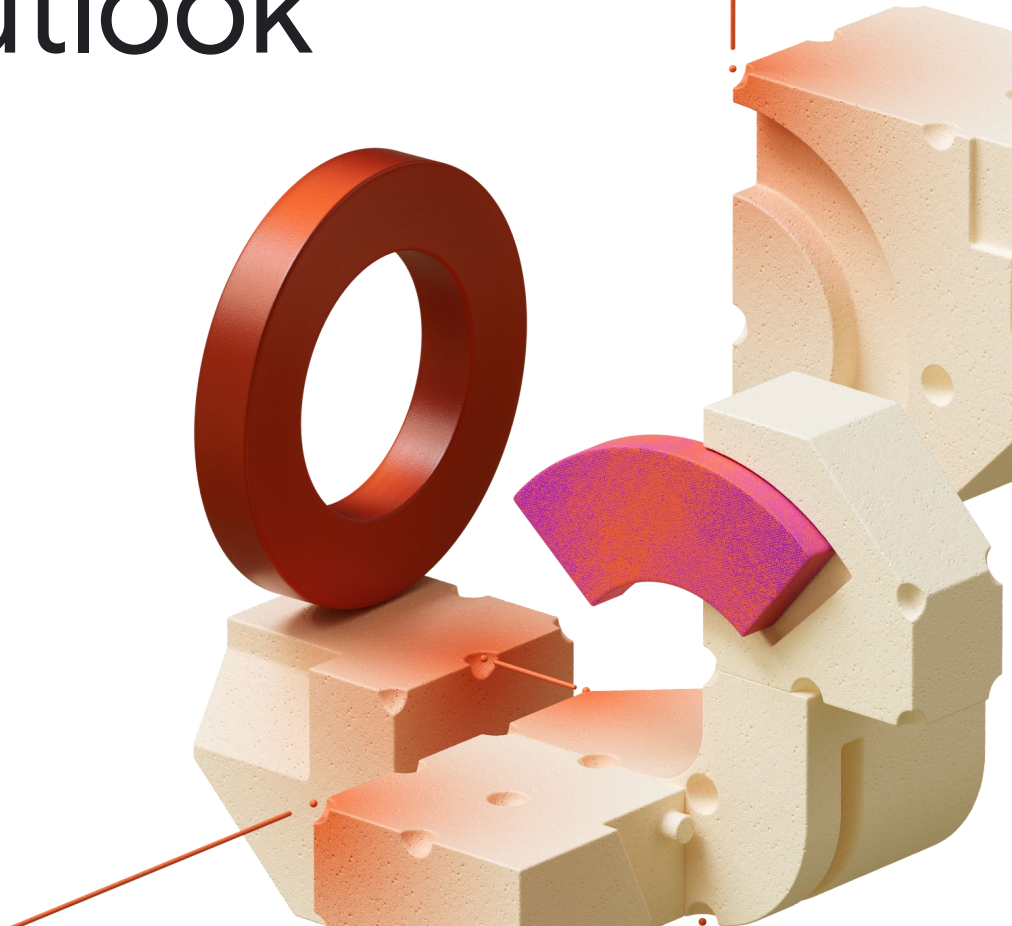


“Slide the pieces of this sliding puzzle around one-at-a-time until all edges align.”



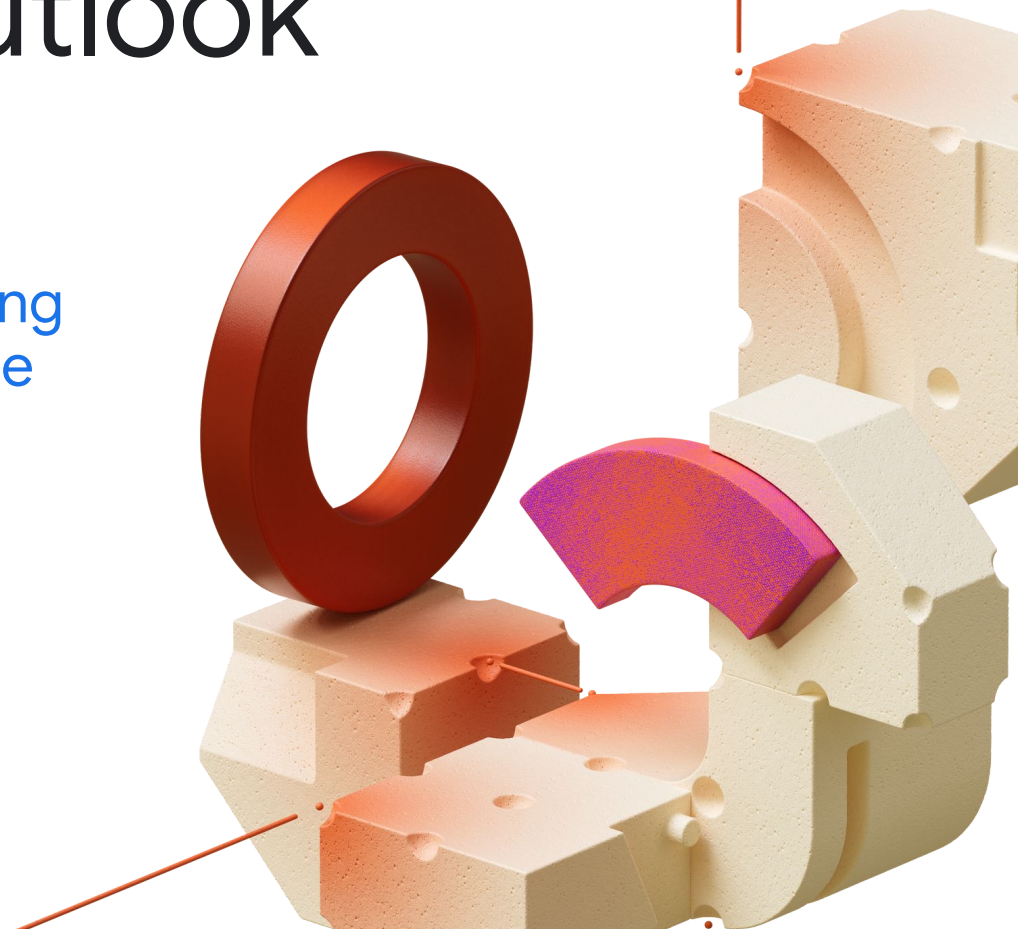
“A hand takes the fitting puzzle piece from the right, rotates it to be in the correct orientation, then puts it into the hole, completing the puzzle. Static scene, no pan, no zoom, no dolly.”

Summary & outlook



Summary & outlook

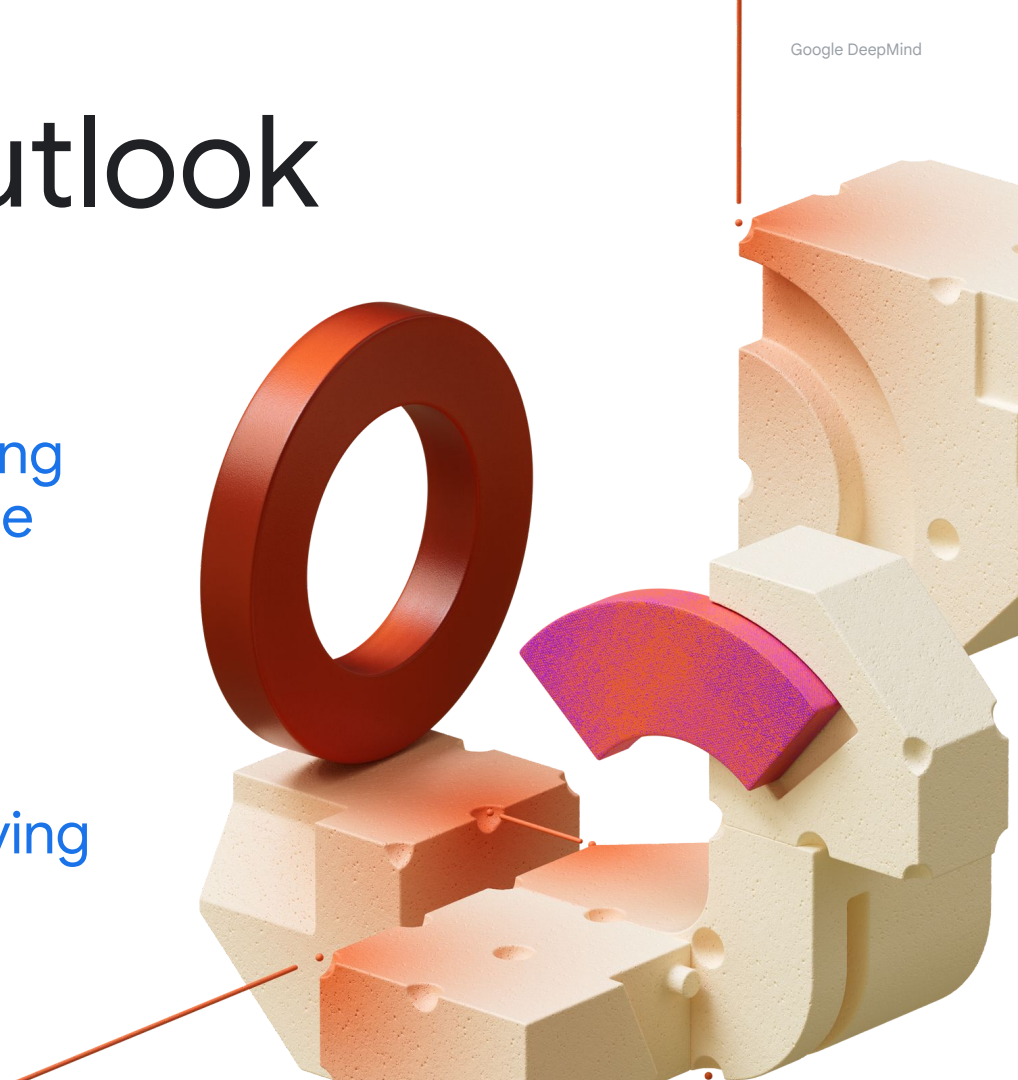
Video models show emergent zero-shot learning and reasoning similar to the early stages of the LLM revolution.



Summary & outlook

Video models show emergent zero-shot learning and reasoning similar to the early stages of the LLM revolution.

If they continue to follow LLMs along a similar trajectory, they might be the path towards solving visual intelligence.



TL;DR

video models

=

visual foundation models

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

Thank you!

And thanks to everyone who supported or enabled this project in various ways:

Oyvind Tafjord • Mike Mozer • Katherine Hermann • Andrew Lampinen • Viorica Patraucean • Shiry Ginosar • Ross Goroshin • Abhijit Ogale • Claire Cui • Kun Zhan • Anish Nangia • Tuan Anh Le • Medhini Narasimhan • Pieter-Jan Kindermans • Shelly Sheynin • David Fleet • Jon Shlens • the Veo Team