ELLIPSIS · LP LEARNING SESSION

WORLD MODELS

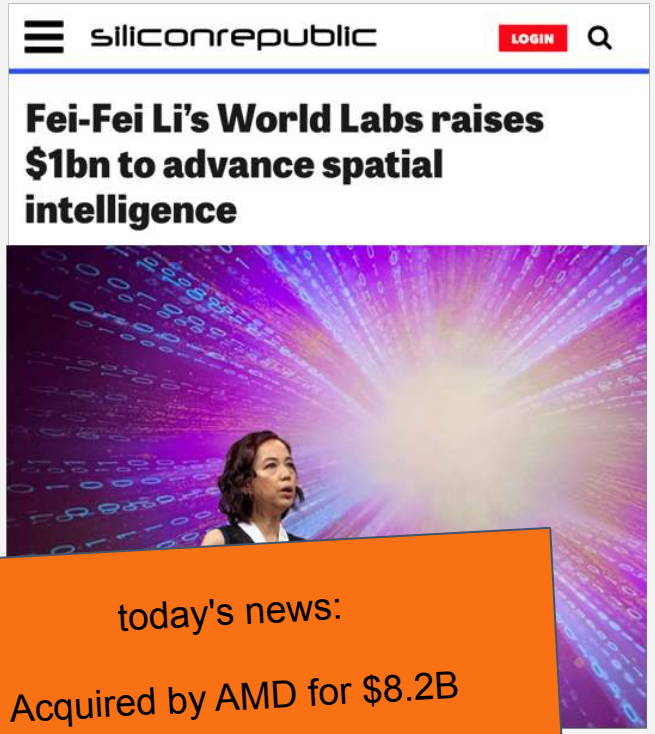
How machines learn to simulate the world and predict the outcomes of actions.



ellipsis

WHAT IS THIS ALL ABOUT

Three headlines from 2026: world-model start-ups raising billions before they have a product.



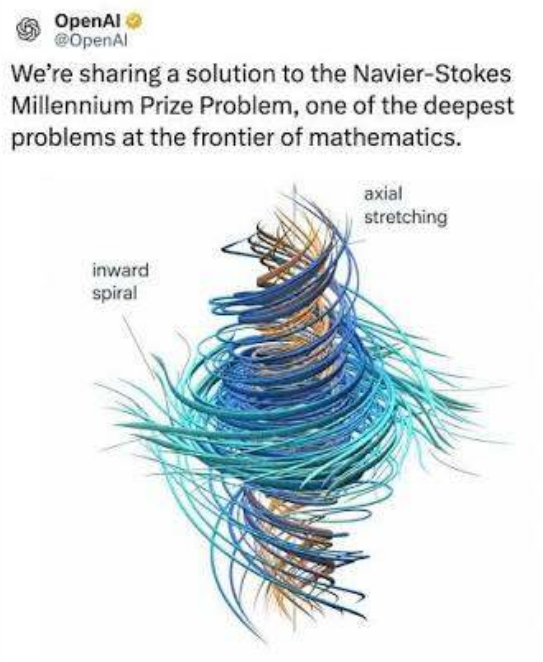




Why are pre-product "World Model" companies snapping up Billions in funding?

THE MORAVEC PARADOX

Computers beat us at maths and chess and cannot kick a ball. The physical world is the hard part.





"how many 'R' are in strawberry?"

Hard for us, easy for machines: maths, chess, code.

Easy for us, hard for machines: catching a ball, folding a shirt.

HOW DO WE LEARN?

How a child learns to kick: told, trying, watching, imagining. AI has done the first three.



being told A little. Eye on the ball.	**2012**	deep learning ImageNet, AlexNet
trying Thousands of kicks. Feedback from the ball, not from us.	**2016**	reinforcement learning AlphaGo
observing A lot. He watched his bigger brother.	**2020**	foundation models GPT-3
imagining He plays the kick in his head before he takes it.	**now**	world models all three, plus imagining

SUPERVISED LEARNING

Being told: people label millions of examples and the model learns to recognise the next one.

LABELLED EXAMPLES



PROCESS

the model learns

NEW PICTURE



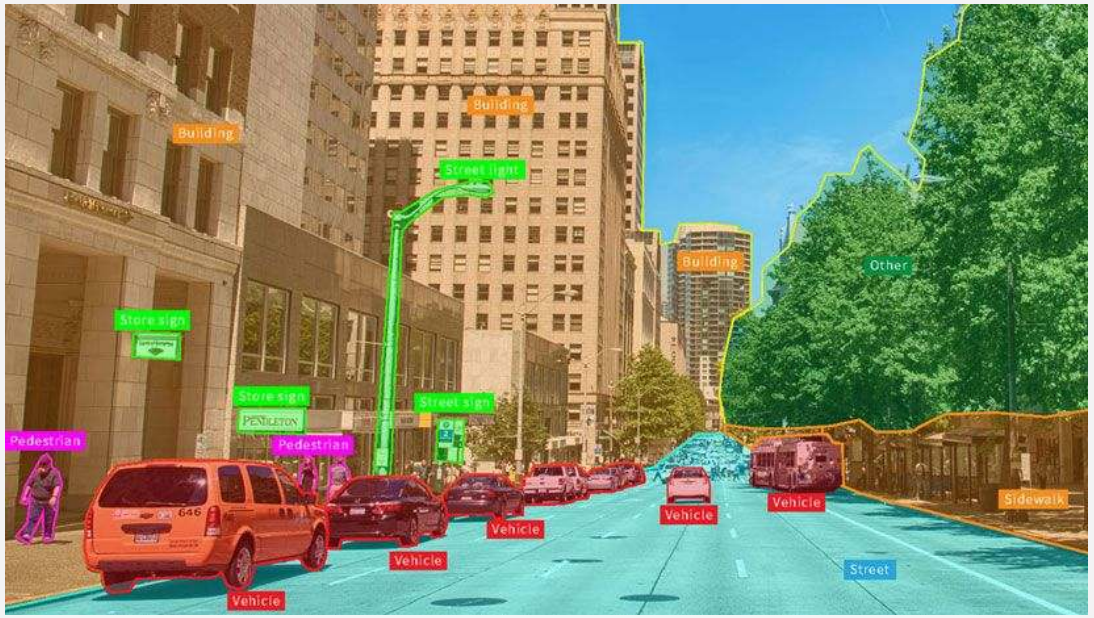
yes, ball

The deep learning era.

ImageNet: 14 million hand-labelled images, 49,000 workers, led by Fei-Fei Li (2009). AlexNet on ImageNet in 2012 started the decade, and my PhD.

LABELING IS HARD TO SCALE

Every label is a person's work. Self-driving needed armies of labellers, and a decade.



24 million boxes

to label six hours of driving (Waymo Open Dataset)

$100B

estimated annotation cost for Waymo

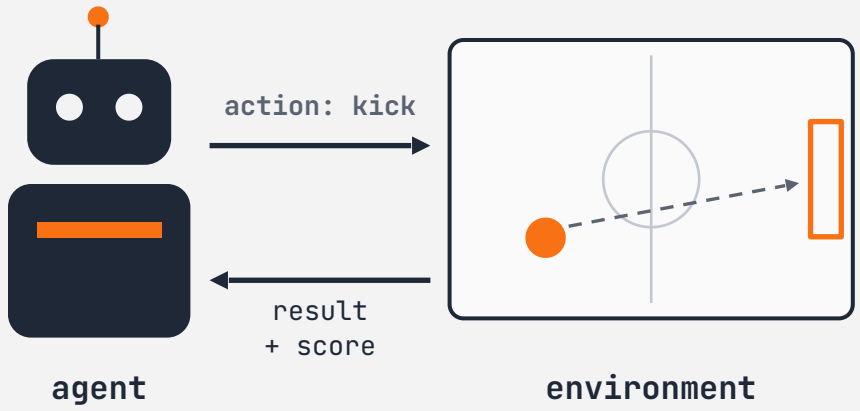


More than a **decade later** driverless service in about a dozen US cities.

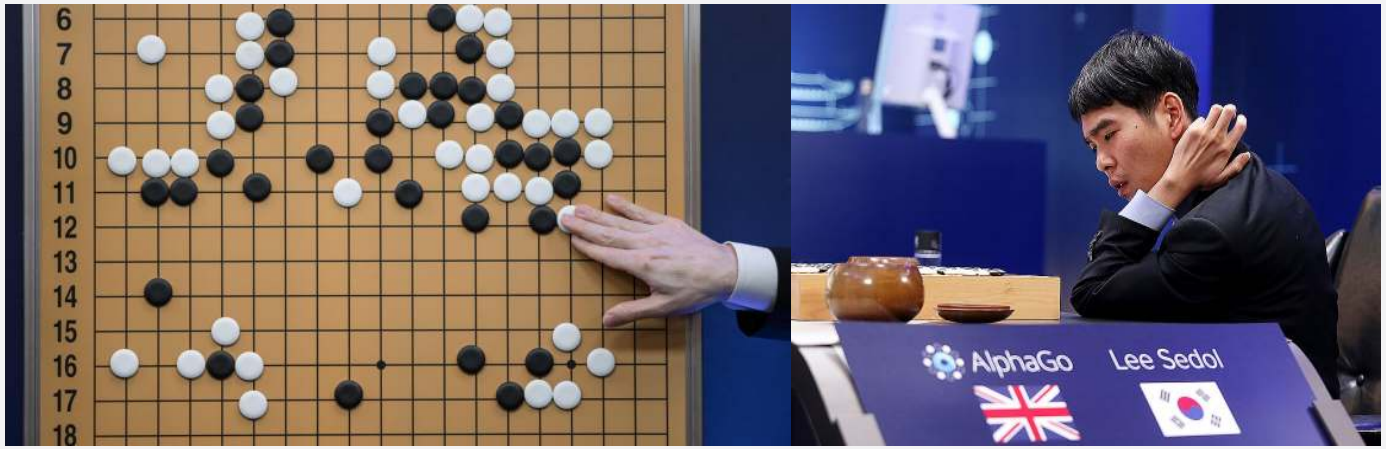
REINFORCEMENT LEARNING

Trying: the model acts, gets a score, keeps what worked. Go was the perfect place for it.

THE IDEA



Kick, see where the ball goes, keep what worked.



2016 · AlphaGo beats Lee Sedol 4–1, after millions of games against itself.

Go has both: a place to try, and a score to keep. Where else do we have both?

WHERE REINFORCEMENT LEARNING WORKS

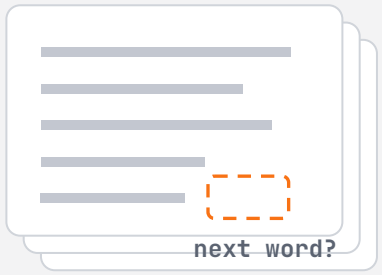
Four things trying needs, across five fields. Digital has them; the physical world does not.

	digital			physical	
● holds ◐ partly ○ missing	**games**	**code and maths**	**writing**	**self-driving**	**robotics**
ENVIRONMENT somewhere to act, reset, try again. Cheap and safe.	● the game itself	● real developer environment	● the text itself	◉ the road, or a hand-built sim	◉ the lab, or a hand-built sim
FIDELITY behaves like the real task (the sim-to-real gap)	● it is the task	● the code really runs	● the text is the product	○ sims miss other drivers, weather	○ sims miss contact, cloth, people
REWARD an automatic score, no human in the loop	● win or lose	● compiles, tests pass	◉ grammar, content is subjective	◉ no crash, progress made	○ what counts as folded?
EXPLORATION random tries score sometimes, so learning can start	● random play is ok	○ random code never runs	○ random words never make a sentence	◉ start from logged driving	○ from scratch, nothing scores

SELF-SUPERVISED LEARNING

Observing: hide part of the data, let the model guess it, check. Nobody labels anything.

TEXT · 2020 · HIDE THE NEXT WORD, GUESS IT



"The highest mountain on Earth is ___"

Mount	62%
Everest	31%
the	3%

it guesses, then the real text tells it if it was right

PICTURES · 2022 · BURY A PHOTO IN NOISE, GUESS IT BACK



prompt

"a boy on the pitch at the Champions League final"

starts from static, ends at a picture nobody took



WHY IT MATTERS

The data is its own answer key.

No labels needed, so it scales to everything ever written or photographed. Llama 3 read 15 trillion words.

- Answers questions it was never trained on.
- Learns a new task from two examples.
- More data, bigger model: predictably better.

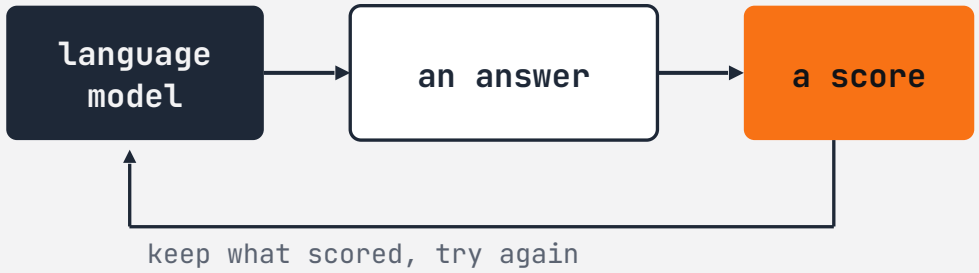
⇒ **Plausible first drafts of anything.**

Answers, code, letters, pictures, video.

GPT-3, 2020 · Stable Diffusion, 2022

PRE-TRAINING + REINFORCEMENT LEARNING

Reading gives the first draft. Trying on top, scored by a test or a person, makes it correct.



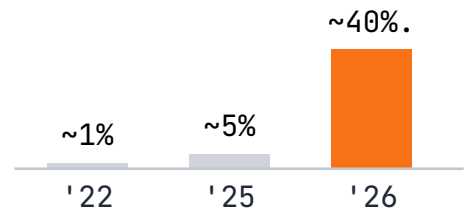
Environment it worked:

- **Chat:** Scored by people (RLHF): people rank the answers. ChatGPT, 2022.
- **Verifiable domains (coding, math):** Scored by a test (RLVR): the code runs, the proof checks.

BIG BANG: CHATGPT

ChatGPT

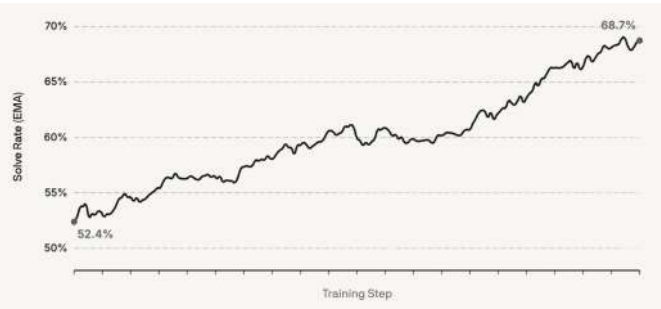
RL SHARE OF TRAINING COMPUTE



CODE · SWE-BENCH

52% → 69%

solve rate during RL



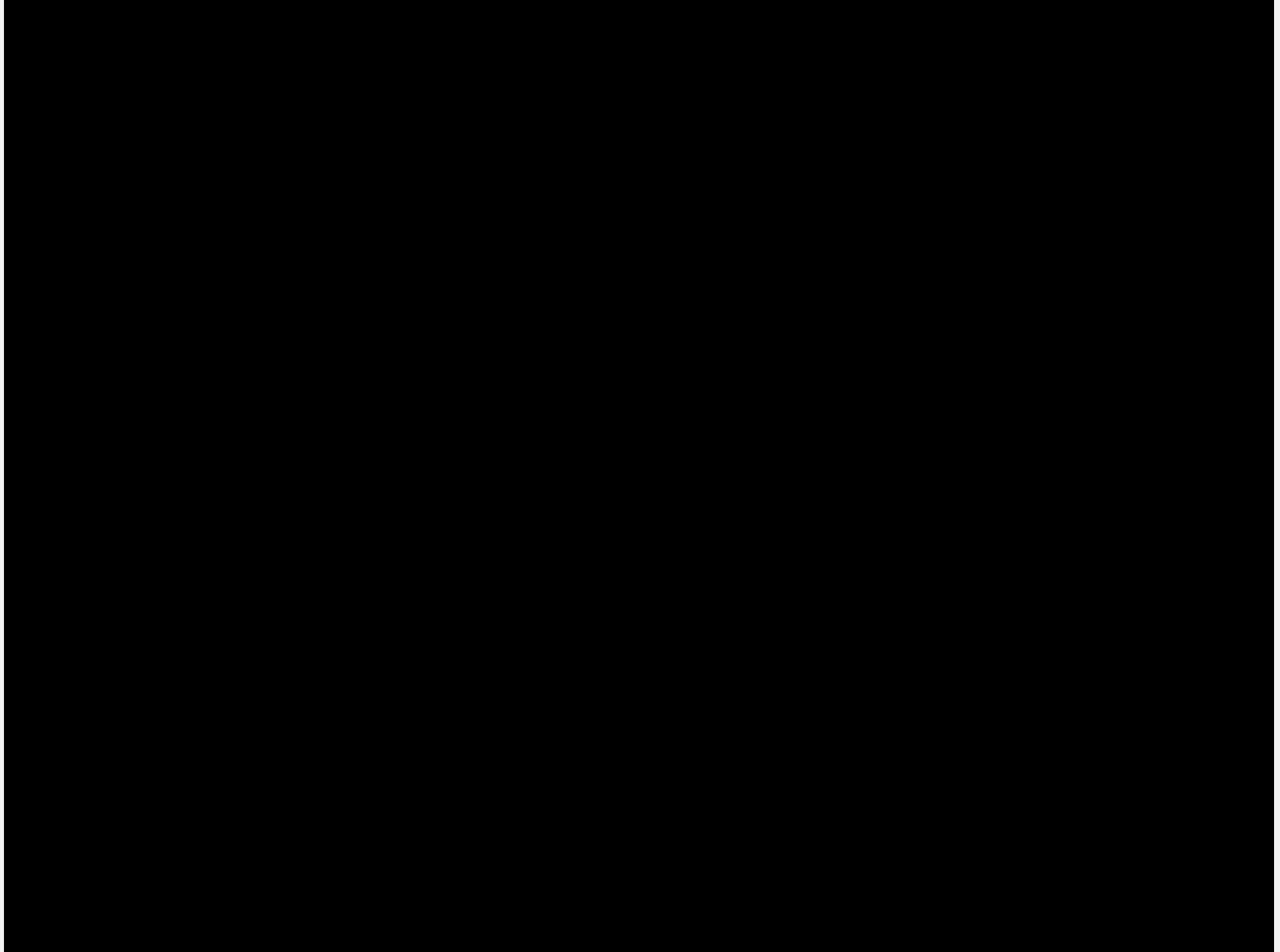
WHAT PRE-TRAINING CHANGED

After pre-training every digital column holds. The physical columns still have nowhere to try.

orange: what changes

● holds ◐ partly ○ missing

	digital			physical	
	games	**code and maths**	**writing**	**self-driving**	**robotics**
ENVIRONMENT somewhere to act, reset, try again. Cheap and safe.	●	● real developer environment	● the text itself	◉ the road, or a hand-built sim	◉ the lab, or a hand-built sim
FIDELITY behaves like the real task (the sim-to-real gap)	●	● the code really runs	● the text is the product	○ sims miss other drivers, weather	○ sims miss contact, cloth, people
REWARD an automatic score, no human in the loop	●	● compiles, tests pass	◉ grammar, content is subjective	◉ no crash, progress made	○ what counts as folded?
EXPLORATION random tries score sometimes, so learning can start	●	● from pre-training	● from pre-training	◉ start from logged driving	○ from scratch, nothing scores





WORLD MODELS INTUITION

The same frame, two actions, two futures generated by a video model.

STATE · one frame



\+

an action

“kick”

“stop”

WHAT HAPPENS NEXT

“kick”



WHAT HAPPENS NEXT

“stop”

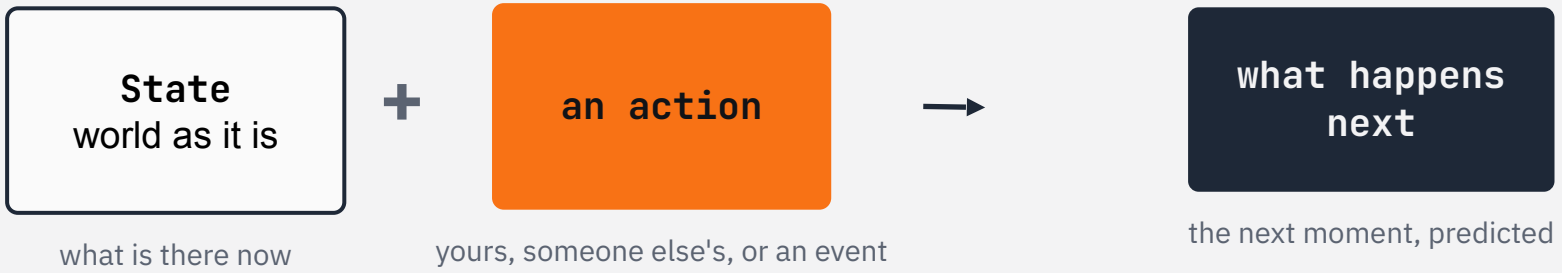


imagining

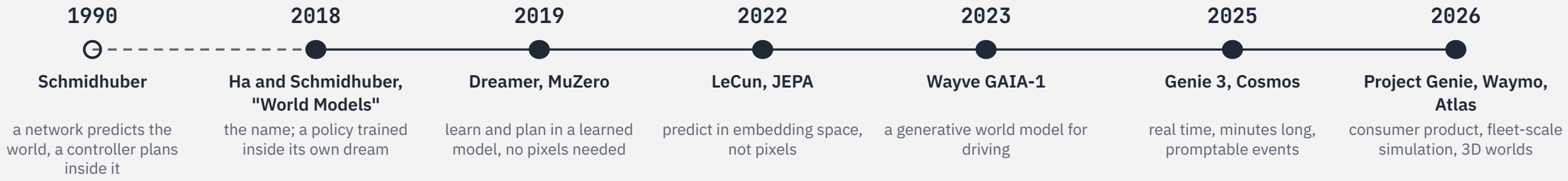
ellipsis

WORLD MODEL: A MODEL OF HOW THE WORLD WORKS

A world model predicts how the world changes when something acts on it.



If a model can predict that, you can try things inside it before you try them in the world.



WORLD MODEL EXAMPLE

One photo went in. Everything you see is generated, frame by frame, from the key presses.



INPUT



one photo. No 3D model, no game engine.

ACTION

← ↑ ↓ →

Each key press goes into the model, which draws the next frame, 24 times a second. Turn around: the columns are still there.

Genie 3

Google DeepMind, Genie 3, August 2025. Real time, 720p, consistent for a few minutes.

SAME MODEL, WITH HANDS

One sentence went in. The keys move the roller, and the paint stays on the wall.



INPUT

“POV action camera of a tan house being painted by a first person agent with a paint roller”

one sentence

ACTION

← ↑ ↓ →

The agent acts in first person. Look away and back: the paint is still there. Nothing was filmed.

Google DeepMind, Genie 3, August 2025. Real time, 720p, consistent for a few minutes.

WHAT WORLD MODELS CHANGE

World models fill the environment and the first tries. Truth when you act, and the score, stay open.

orange: what changes

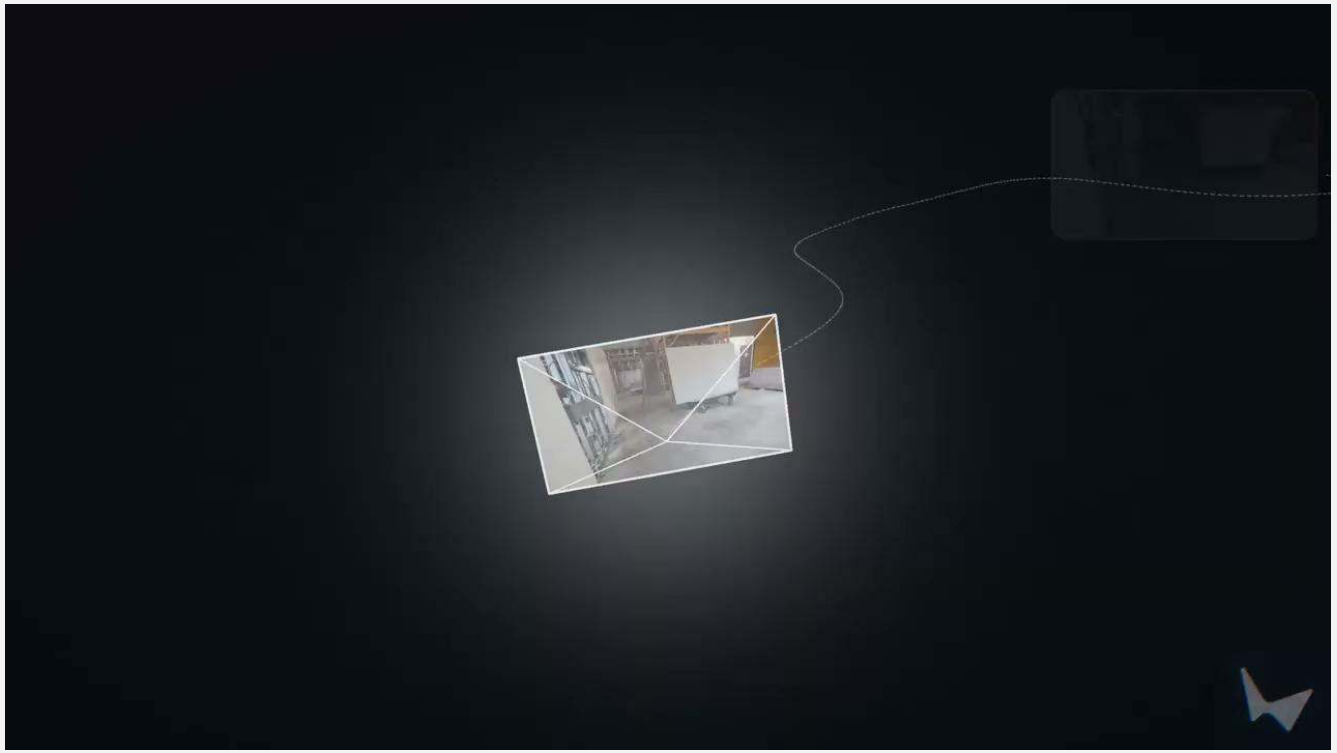
● holds ◐ partly ○ missing

	digital			physical	
	games	coding	writing	self-driving	robotics
ENVIRONMENT somewhere to act, reset, try again. Cheap and safe.	●	●	●	● learned from fleet video (Waymo)	● learned from video (Genie, GWM-1)
FIDELITY behaves like the real task (the sim-to-real gap)	●	●	●	◉ true when you act?	◉ true when you act? contact still weak
REWARD an automatic score, no human in the loop	●	●	●	◉ no crash, progress made	◉ a judge model, weak
EXPLORATION random tries score sometimes, so learning can start	●	●	●	● pre-trained agents	● pre-trained agents

USE 1 · ROBOTS, THE GYM

A phone video of a room becomes a simulation. The robot practises there, not in the lab.

THE GYM the world + an action → what happens next



Video: World Labs Atlas, Sep 2026. Evaluation: Runway Research, 8 policies, 1,450 rollouts.

HOW IT WORKS

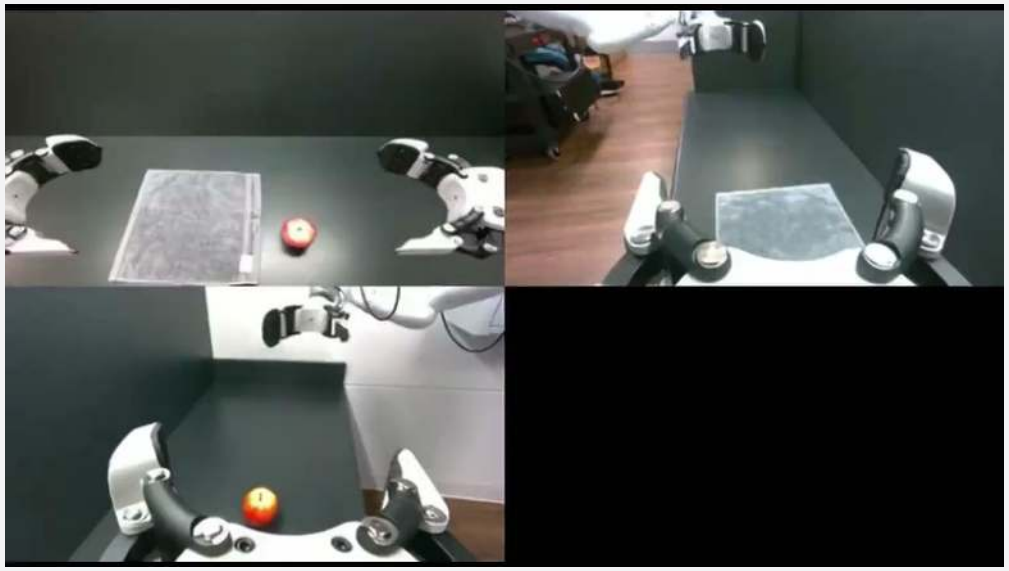
- 24 frames from a phone become a 3D room the robot can move in.
- It renders what the robot's cameras would see.
- Change the objects, the light, the layout: thousands of tries overnight.

USE 1 · ROBOTS, THE DRIVER

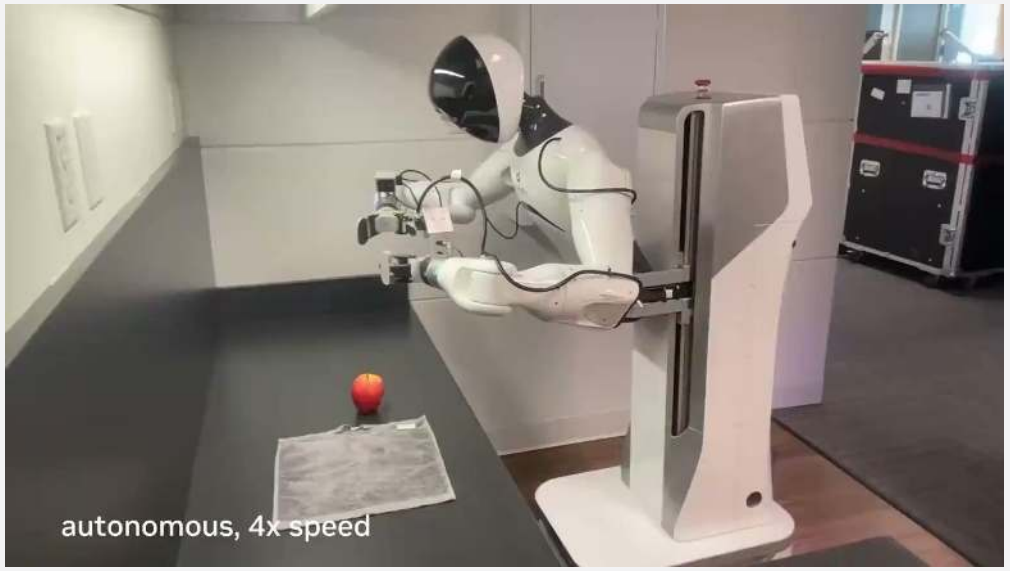
The model imagines the task as a video, then reads the robot's movement off it.

THE DRIVER the world + a goal → the action · a world action model

GENERATED SENSOR DATA · three cameras



RECORDING OF THE REAL ACTION



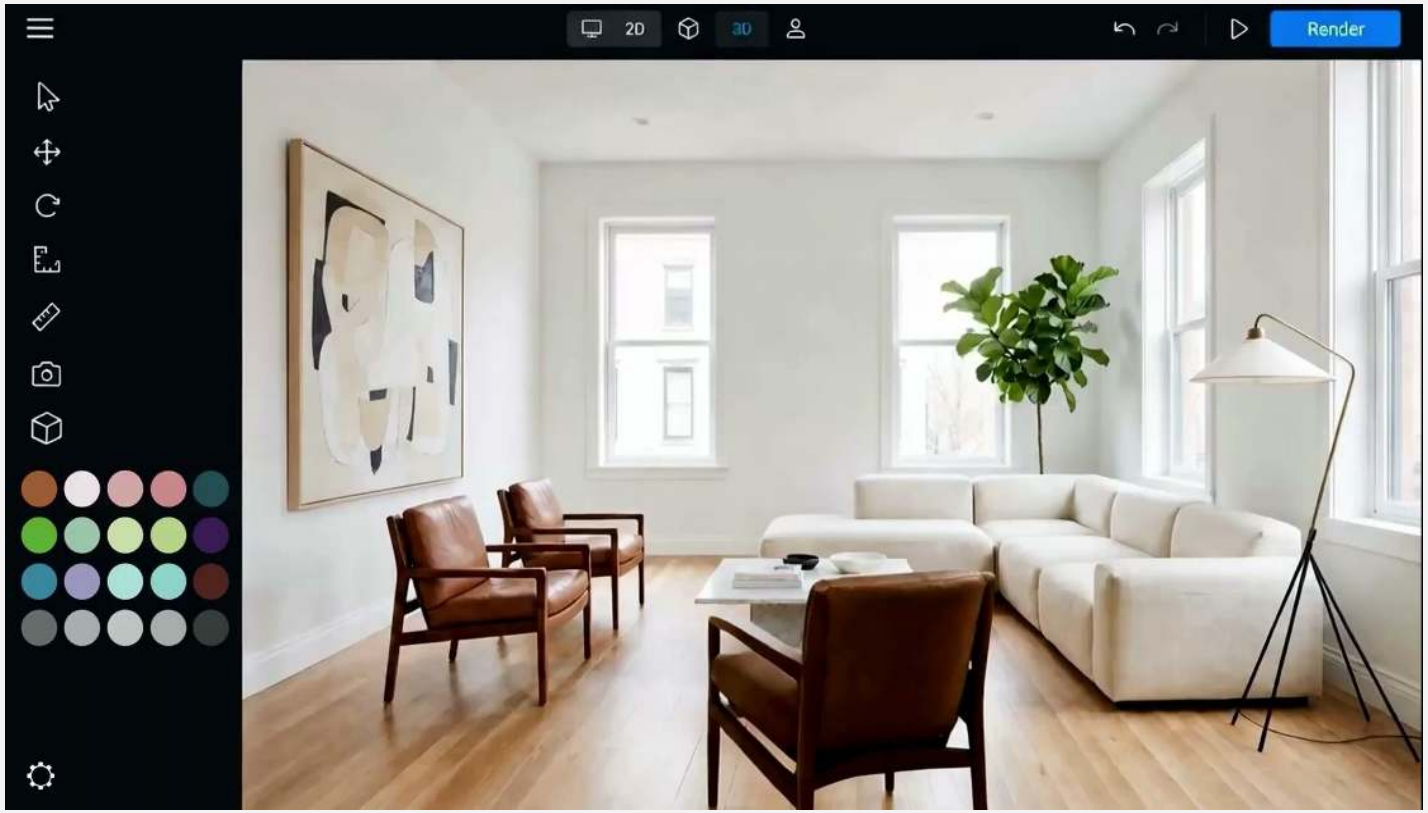
7 Hz
control loop, from a 14B video model

2×
better on unseen tasks than today's robot policies

30 min
of play data to adapt to a new robot

USE 2 · SOFTWARE WITHOUT CODE

The model takes your click and draws the next screen. No code in between.



- Clicks in, pixels out, in real time
- No HTML behind it
- Prototypes, adaptive interfaces, lessons
- A gym for computer-use agents
- Runway Solaris, 31 Aug 2026

USE 2 · SOFTWARE WITHOUT CODE

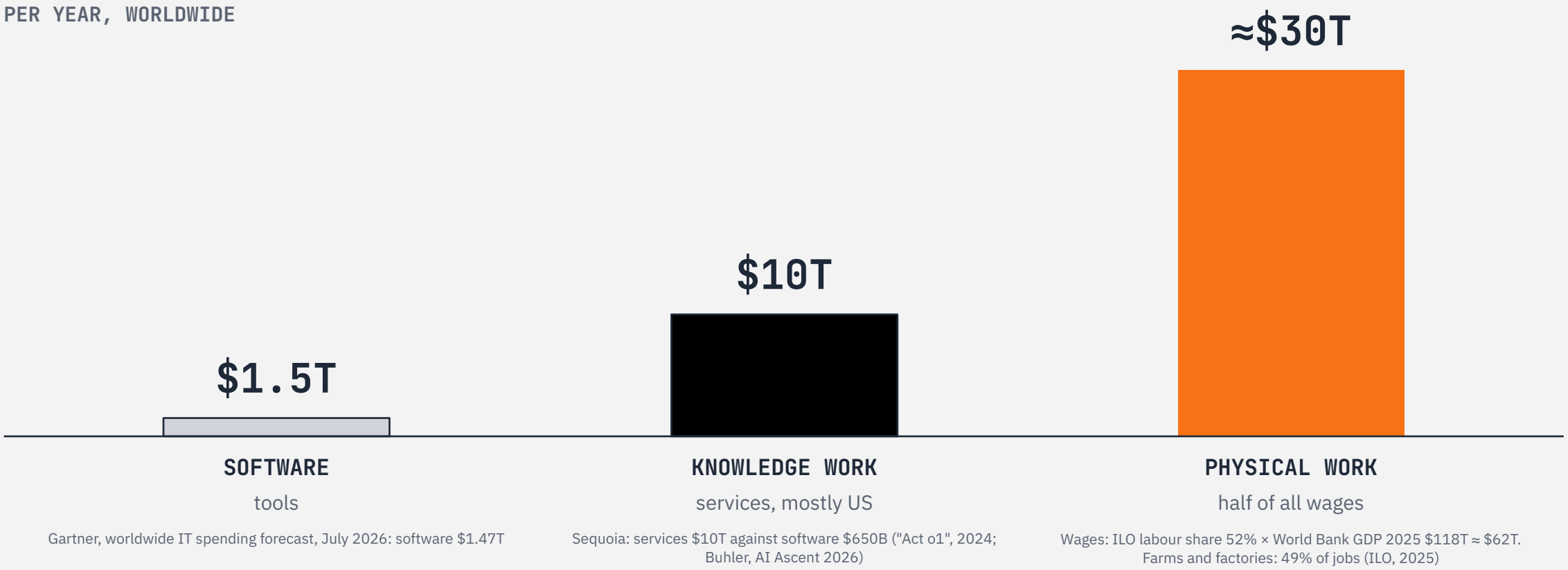
The model takes your click and draws the next screen. No code in between.



- Clicks in, pixels out, in real time
- No HTML behind it
- Prototypes, adaptive interfaces, lessons
- A gym for computer-use agents
- Runway Solaris, 31 Aug 2026

WHY IT MATTERS

AI stops selling tools and starts selling work. The biggest pool of work is physical.



Gartner, worldwide IT spending forecast, July 2026: software $1.47T

Sequoia: services $10T against software $650B ("Act o1", 2024; Buhler, AI Ascent 2026)

Wages: ILO labour share 52% × World Bank GDP 2025 $118T ≈ $62T. Farms and factories: 49% of jobs (ILO, 2025)

AI already sells knowledge work: a quarter to a third of the code at Microsoft and Google. Physical work: nowhere to try.

THE MONEY AND THE MODELS

The West, Google aside, has no competitive video model, and bets billions on what sits on top.

MEGA ROUNDS · WORLD AND VIDEO MODELS, LAST 12 MONTHS

company	what it builds	round	valuation	when
AMI Labs	JEPA, no pixels	**$1.03B**	$3.5B	Mar 26
World Labs	3D worlds: Marble, Atlas	**$1B**	$5.4B	Feb 26
Luma	video model, Ray 3	**$900M**	≈$4B	Nov 25
Emulate	ex-Genie team, in talks	**≈$700M**	≈$3.5B	Sep 26
GigaAI	driving and robot world models	**≈$490M**	—	Mar–Jun 26
General Intuition	agents from game clips	**$320M**	$2.3B	Jun 26
Runway	video model plus world model	**$315M**	$5.3B	Feb 26
Odyssey	interactive worlds	**$310M**	$1.45B	Jun 26
Decart	real-time video; Anthropic in talks	**$300M**	$4B	May 26

plus $1.44B disclosed by Chinese world-model start-ups by Aug 2026 (TechBuzz China); Forbes: over $3B in H1 2026

BEST VIDEO MODELS · ARENA, SEP 2026

#	model	maker	elo
1	**Gemini Omni Flash**	Google	1233
2	**Wan 3.0**	Alibaba	1229
3	**MiniMax H3**	MiniMax	1227
5	**Seedance 2.0**	ByteDance	1210
6	**MAGI-2**	Sand.ai	1156
8	**HappyHorse 1.1**	Alibaba	1147
11	**SkyReels V4**	Skywork	1095
12	**Kling 3.0**	Kuaishou	1095
14	**Veo 3.1**	Google	1088

Text-to-video, one entry per model family; rank as listed. Not in the top 15 of either board: OpenAI, Runway, Luma, Pika. First US start-up on image-to-video: xAI, 8th. Lightricks (IL) 22nd. MiniMax H3 ships its weights.

CURRENT LIMITATIONS AND LINES OF WORK

Not true enough when you act, no score, and years behind text on infrastructure.

FIDELITY

- Physics is approximate. Memory lasts a few minutes.
- Trained on successes: it predicts the usual, not the consequence of your action.
- No touch: video has no force, weight or friction. Hands, cloth and liquids fail.

INFRA

- Up to $0.40 a second of video;
- Real time only for small models at low resolution.

OPEN WORK

- Predict state, not pixels: latent diffusion, JEPA.
- Actions from unlabelled video: latent actions.
- Long memory through 3D. Engine plus video hybrids.
- Real-time distillation: few-step, causal models.
- A model that scores the task: learned rewards.
- Hierarchical planning inside the model.

WHERE WILL THE VALUE BE CAPTURED?

Three ways to invest in world models.

1

GENERAL MODELS

The next foundational World Model

General world models for any place and any agent. A handful of labs, billion-dollar rounds, Google and NVIDIA as competitors.

2

NEW MODALITIES

Worlds that video does not show.

Chemistry, cell biology, materials, quantum, plasma, the grid. Data nobody has pre-trained on, and value no general model will capture.

3

ASSUME IT GETS SOLVED

Picks and shovels, and what is built on top.

Action data, evaluation and scoring, simulation tooling, managing the models in production, and the applications that use them.


ellipsis