



World Models: Video, Physics and Planning

Frontier research 01 | What a world model must do beyond generating video, the four families as of 2026 (latent control models, interactive video, video foundation models, JEPA), compounding rollout error computed, how to evaluate physics and usefulness, open questions

#WorldModels #VideoModels #ModelBasedRL #2026

What a world model is, and is not

World model

In one sentence
●○○○

A model that can answer: if I do this, what happens next?

Practitioner
●●○○

A learned simulator: state + action → next state, reward, done. Agents plan or train inside it.

- cheaper and safer than real trials
- only as good as its predictions

Expert
●●●○

Two design choices dominate: predict pixels or a latent state, and learn actions from labelled control data or infer them from video.

- pixels: watchable, expensive, wasteful
- latent: compact, hard to inspect

Under the hood
●●●●

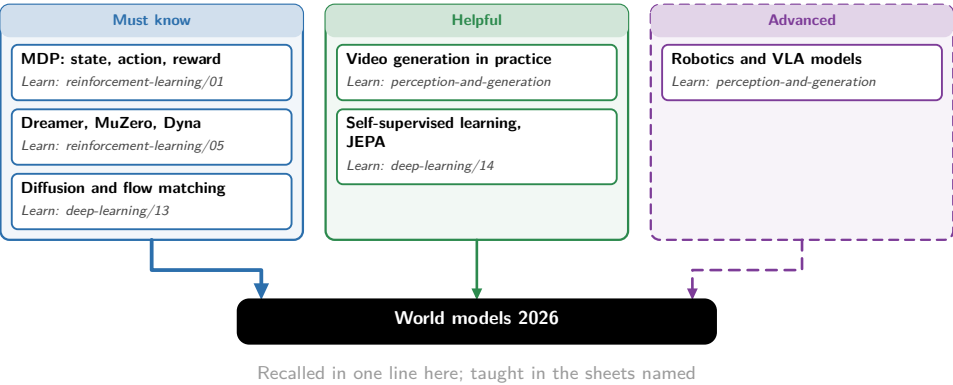
Open question: does training to predict the next frame produce a model of causes (physics, objects, agents) or a model of typical-looking video?

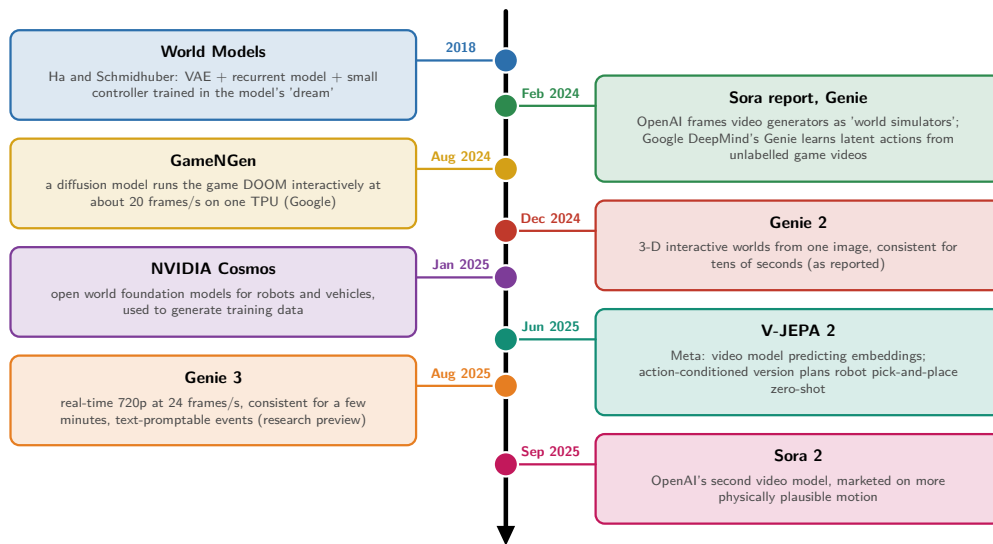
- realism ≠ correctness
- evaluation is the bottleneck

Analogy

A flight simulator is useful to a pilot only if the aircraft in it reacts like the real one, however good the scenery looks.

One concept at four depths. The last level is where the 2026 research questions sit



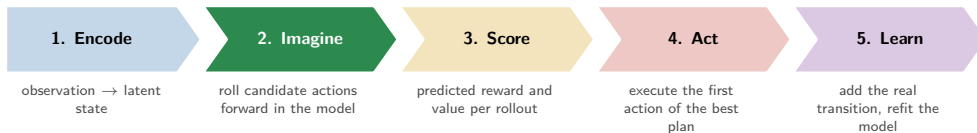


Milestones after Dreamer and MuZero (public announcements; model-based RL history is in reinforcement-learning/05)

⚠ What a demo does not show

A curated clip proves the model can produce that clip, not how often. Ask for the share of rollouts that stay consistent, the horizon in seconds, and whether an agent trained inside the model succeeds in the real environment.

© Using a world model to act



The loop of a model-based agent. Only the last step touches the real world; every other step costs compute, not risk

Error after an H-step rollout (bound for a Lipschitz model)

FORMULA $E = e \times (L^H - 1) / (L - 1)$

WHERE

- E = worst-case state error after H predicted steps (state units)
- e = one-step prediction error of the model (state units)
- L = Lipschitz constant of the dynamics: how much a small state error can grow in one step (> 1) (ratio)
- H = rollout horizon (steps)

WORKED

e = 0.01
L = 1.1
H = 15
E = 0.01 × (1.1¹⁵ - 1) / (1.1 - 1)
E = 0.3177

RESULT E = 0.3177

READ IT A 1 % one-step error becomes about 32 % after 15 steps when errors grow 10 % per step. This is why Dreamer-style agents imagine only short horizons (about 15 steps in DreamerV1) and bootstrap the rest with a value function.

MISTAKES

- Reporting one-step accuracy as if it held over long rollouts
- Using long imagined horizons without a value estimate at the end

PROBLEM A model must generate 1 minute of play at 24 frames/s. How much time per frame, and how likely is the minute to stay consistent if each frame independently keeps the scene consistent with probability 0.999?

GIVEN

- f = 24 frames per second, T = 60 s
- q = 0.999 per-frame consistency (assumption for the example)

1 Frames to generate

$$N = f \times T = 24 \times 60 = 1440$$

Why: each frame is predicted from the previous ones

2 Time budget per frame

$$1000 \text{ ms} / 24 = 41.7 \text{ ms}$$

Why: the whole denoising pass must fit in one frame period

3 Consistency over the minute

$$0.999^{1440} = e^{(1440 \times \ln 0.999)} = e^{(-1.44)} = 0.237$$

Why: errors compound frame after frame

4 Per-frame rate for 90 %

$$q = 0.9^{(1/1440)} = 0.99993$$

Why: invert $q^N = 0.9$

ANSWER 41.7 ms per frame; only about 24 % of minutes stay consistent at q = 0.999; 90 % needs failures cut by a factor of about 14

CHECK ✓ $0.99993^{1440} = e^{(-0.1008)} \approx 0.904$

The real-time budget of an interactive video world model (illustrative consistency figure, not a measured one)

✓ Evaluating a world model

Axis	Question	Typical measure (as of 2026)
Visual quality	Does it look real?	FVD, human preference; says nothing about physics
Physical plausibility	Do objects fall, collide, flow correctly?	Physics-IQ, VideoPhy-style benchmarks: predict real filmed outcomes
Controllability	Does each action produce its expected effect?	action-following accuracy, inverse-dynamics checks
Consistency	Is a place the same when you come back?	revisit tests, object permanence, horizon before drift
Downstream use	Does an agent trained or planning in it succeed for real?	success in the real environment or a held-out simulator

What to measure, from weakest to strongest evidence of a usable world model

- **Realism is not physics** Physics-IQ (Motamed et al., 2025) reported that physical understanding of leading video models was severely limited and unrelated to visual realism
- **Out-of-distribution** Kang et al. (2024) trained video models on simple simulated physics and reported case-based imitation of training examples rather than learned rules when tested outside the training range
- **Usefulness** the strongest evidence remains control results: Dreamer-style agents and V-JEPA 2 robot planning measure success on real tasks, not frame quality

MYTH	FACT
✗ Scaling video data alone will produce physics	✓ Unresolved, and current evidence is mixed Realism improves with scale; out-of-distribution physical prediction has not followed at the same pace (as of 2026).
✗ A world model must predict pixels	✓ Many predict latents or embeddings MuZero predicts only reward, value and policy; JEPA predicts representations.
✗ Interactive generated games replace game engines	✓ They drift and forget state over minutes Engines keep exact state; generated worlds keep a statistical impression of it.
✗ A good one-step predictor is a good planner	✓ Errors compound over the planning horizon See the bound above: 1 % per step, 32 % after 15 steps.

🚩 Open questions and recap

Open question	Why it matters	Current approaches
Learning causes, not appearances	plans fail where physics is wrong	physics benchmarks, simulated training data, 3-D structure in the model
Long-horizon consistency and memory	worlds drift after minutes	explicit memory of past frames, 3-D scene state, longer context
Actions from unlabelled video	internet video has no action labels	latent action models (Genie), inverse dynamics

Pixels or latents?	pixel models waste capacity on texture	JEPA-style prediction, hybrid decoders
Evaluation that predicts usefulness	realism scores mislead	downstream agent success, revisit tests
Sim-to-real from generated worlds	robots must transfer	domain randomisation, real-data fine-tuning (Cosmos-style)

Open questions on world models (as of 2026) and how they are attacked

Remember

- 01 A world model takes actions
no action input, no planning
- 02 Errors compound over rollouts
keep horizons short, bootstrap with values
- 03 Realism ≠ physics
measure them separately
- 04 Usefulness is the real test
does an agent succeed in the real world?

Q1 What separates a world model from a video generator? A An action input: different actions must produce different predicted futures.
Q2 Rollout error bound with $\epsilon = 0.01$, $L = 1.1$, $H = 15$? A $0.01 \times (1.1^{15} - 1) / 0.1 \approx 0.32$.
Q3 Why does Dreamer imagine only about 15 steps? A Model errors compound; a learned value function estimates what lies beyond the horizon.
Q4 What did Genie add to interactive video models? A Latent actions learned from unlabelled game videos, so a world can be controlled without action labels.
Q5 What does V-JEPA 2 predict? A Embeddings of future video, not pixels; an action-conditioned version plans robot manipulation.
Q6 Frame budget at 24 frames/s? A $1000 / 24 \approx 41.7$ ms per frame.
Q7 What does FVD measure, and what not? A Visual similarity to real video; not physical correctness or controllability.

Q8 What did Physics-IQ (2025) report?

A Physical understanding of leading video models was severely limited and unrelated to visual realism.

Q9 Strongest evidence that a world model is useful?

A An agent trained or planning in it succeeds in the real environment.