



From Perception to Simulation

The Emergence of World Models in Multi-modal Reasoning

Yujun Cai



ANT
GROUP



DeepMind

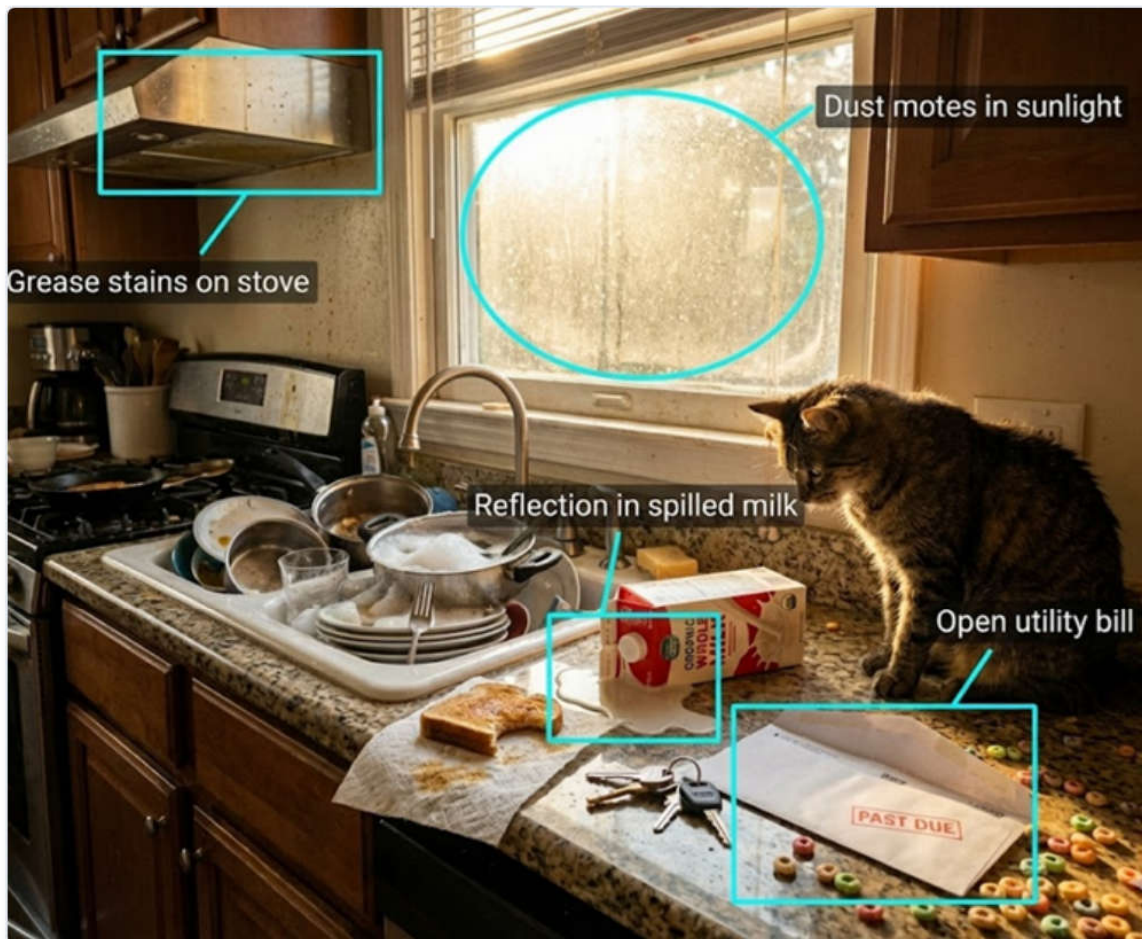


From Perception to Simulation



The central shift: from recognizing the world to internally simulating how it can change.

The First Wave: Perception



PERCEPTION OUTPUTS

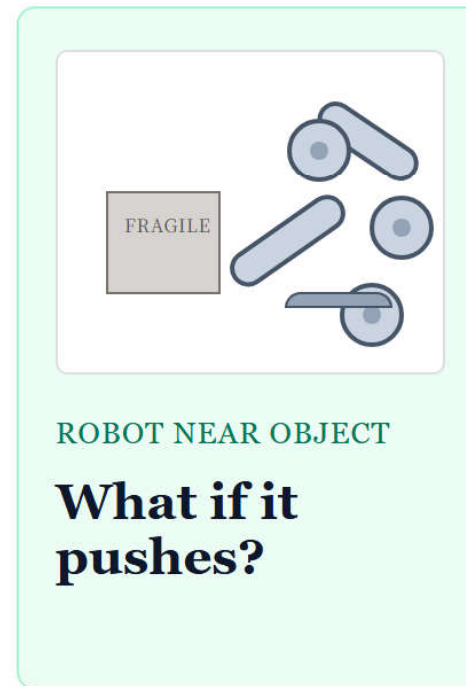
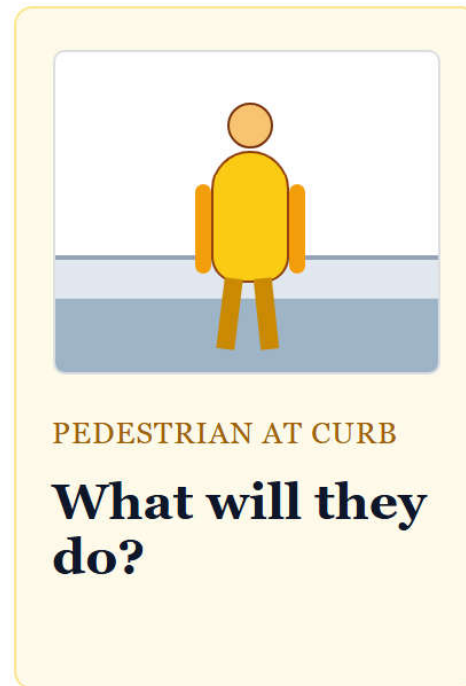
CLASSIFICATION
category

DETECTION
boxes

CAPTIONING
sentences

VQA
answers

Recognition Is Not Enough



Labels identify entities. Reasoning needs dynamics, affordances, and consequences.

From Recognition to Simulation

Recognition → Temporal Reasoning → Prediction → Counterfactuals → Action

The field is moving from static descriptions of the current state toward models of transitions.

From Recognition to Simulation

1

What is it?

recognition

Recognition → Temporal Reasoning → Prediction → Counterfactuals → Action

The field is moving from static descriptions of the current state toward models of transitions.

From Recognition to Simulation

1

What is it?

recognition

2

**What is
happening?**

temporal reasoning

Recognition → Temporal Reasoning → Prediction → Counterfactuals → Action

The field is moving from static descriptions of the current state toward models of transitions.

From Recognition to Simulation

1

What is it?

recognition

2

**What is
happening?**

temporal reasoning

3

**What will
happen?**

prediction

Recognition → Temporal Reasoning → Prediction → Counterfactuals → Action

The field is moving from static descriptions of the current state toward models of transitions.

From Recognition to Simulation

1

What is it?

recognition

2

**What is
happening?**

temporal reasoning

3

**What will
happen?**

prediction

4

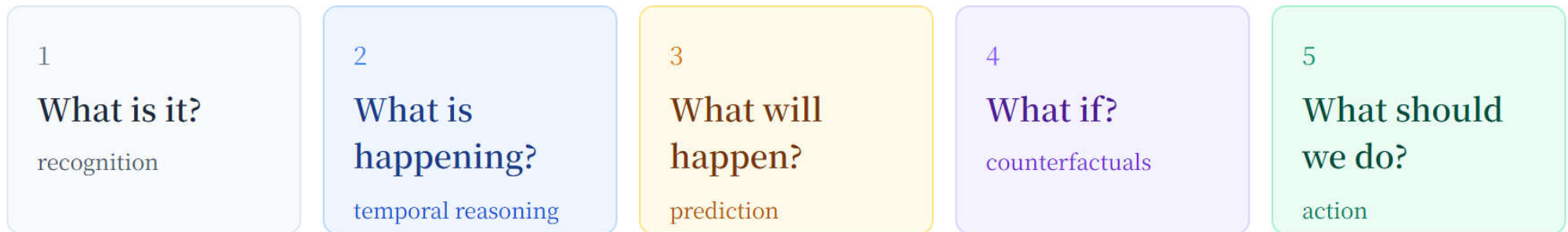
What if?

counterfactuals

Recognition → Temporal Reasoning → Prediction → Counterfactuals → Action

The field is moving from static descriptions of the current state toward models of transitions.

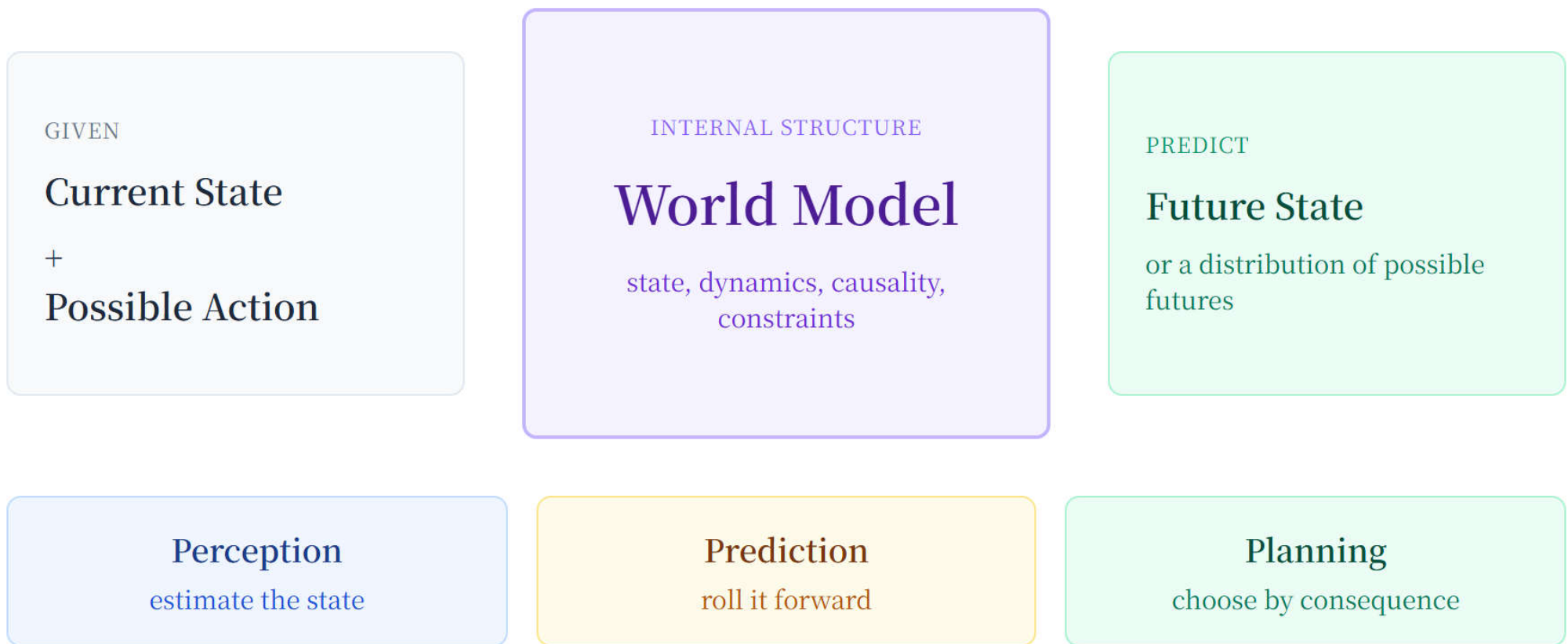
From Recognition to Simulation



Recognition → Temporal Reasoning → Prediction → Counterfactuals → Action

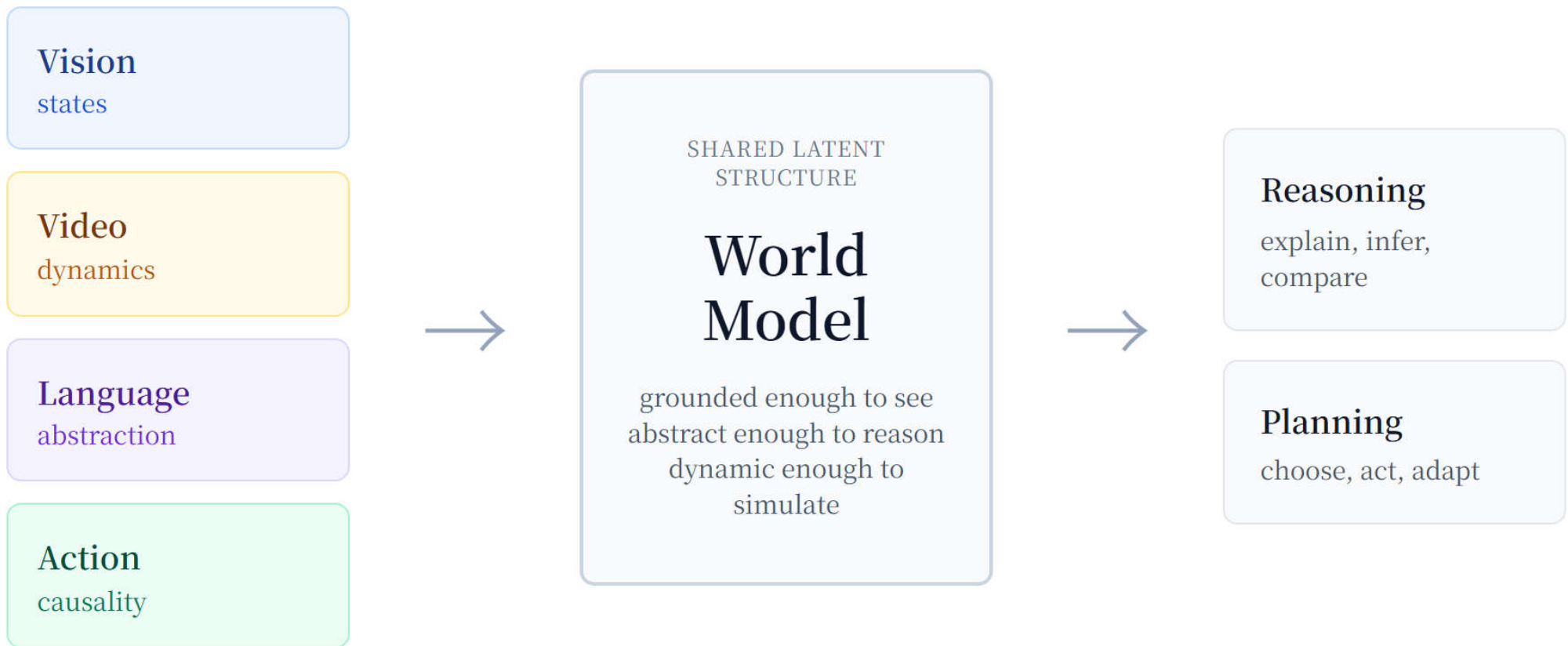
The field is moving from static descriptions of the current state toward models of transitions.

World Models as the Bridge



A world model is not just a fancy term. It is the intermediate structure that connects seeing to deciding.

Why Multimodality Matters



What We Will Explore Today

STEP 1

Represent

state, geometry,
agents, context

STEP 2

Predict

change,
dynamics,
uncertainty

STEP 3

Simulate

interactive and
physical futures

STEP 4

Evaluate

consistency,
commonsense,
limits

STEP 5

Reason

spatial and
causal
understanding

Tutorial map

We will move from world state and dynamics toward simulation, evaluation, and reasoning.

Opening Takeaway

PAST

Describe the scene

recognize objects, attributes, relations

NOW

Predict the state

infer hidden structure and future change

FUTURE

Model the change

learn dynamics, constraints, causal effects

World models turn multimodal perception into a basis for state prediction, physical simulation, and reasoning over possible futures.

Organizers & Speakers



Yujun Cai
@Ant Group& @UQ



Jianfei Cai
@Monash



Yiwei Wang
@UCM



Ming-Hsuan Yang
@UCM



Kai-Wei Chang
@UCLA



Ming-Yu Liu
@NVIDIA



Hang Qi
@Google DeepMind



Dan Kondratyuk
@rekursiv.ai

World Model Tutorial Schedule

Time	Session	Speaker
14:00 - 14:10	Opening Remark: Motivation and Overview [Abstract] [Slides]	Yujun Cai
14:10 - 14:40	Invited Talk: From Chain-of-Thought to Chain-of-State: Why Capable Models Must Predict the World Back [Abstract] [Slides]	Dan Kondratyuk
14:40 - 15:10	Invited Talk: Genie 3: Generating interactive photorealistic worlds [Abstract] [Slides]	Hang Qi
15:10 - 15:50	Invited Talk: Towards Physics-Consistent Efficient Visual World Models [Abstract] [Slides]	Jianfei Cai
15:50 - 16:20	Invited Talk: VideoPhy: Physical Commonsense Evaluation in Video Generation [Abstract] [Slides]	Kai-Wei Chang
16:20 - 16:50	Invited Talk: Cosmos 3: Omni World Foundation Models for Physical AI [Abstract] [Slides]	Ming-Yu Liu
16:50 - 17:25	Invited Talk: Toward World Models: Geometry, View Synthesis, and Visual Reasoning [Abstract] [Slides]	Ming-Hsuan Yang

From Chain-of-Thought to Chain-of-State: Why Capable Models Must Predict the World Back

Dan Kondratyuk