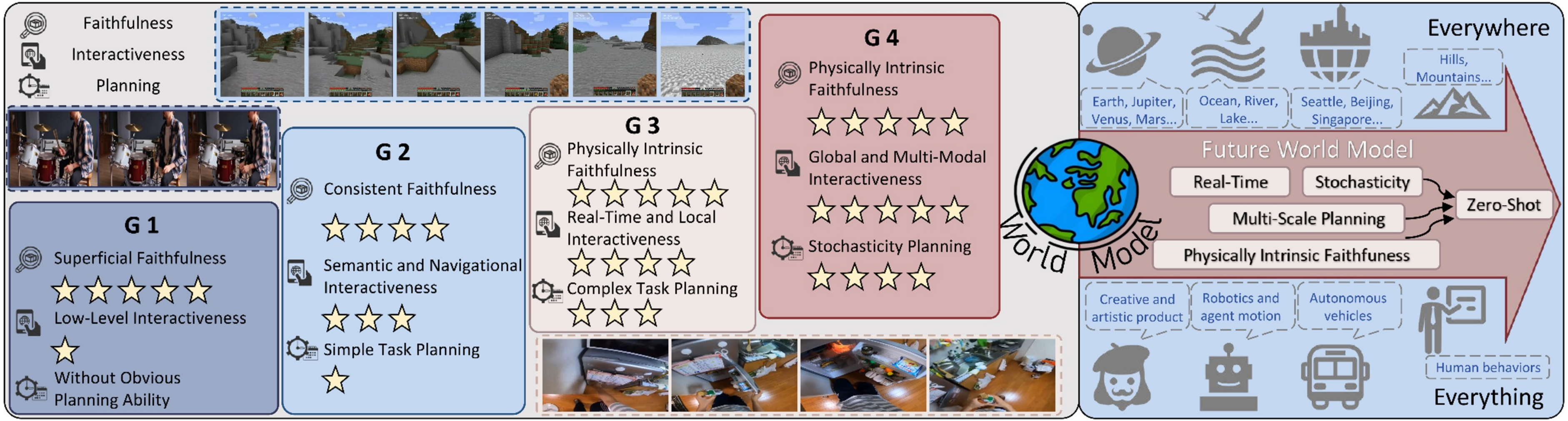
Predictive World Models: Faithfulness, Interactiveness and Planning



Faithfulness

Interactiveness

Planning

G 1

Superficial Faithfulness

☆☆☆☆☆

Low-Level Interactiveness

☆

Without Obvious Planning Ability

G 2

Consistent Faithfulness

☆☆☆☆

Semantic and Navigational Interactiveness

☆☆☆

Simple Task Planning

☆

G 3

Physically Intrinsic Faithfulness

☆☆☆☆☆

Real-Time and Local Interactiveness

☆☆☆☆

Complex Task Planning

☆☆☆

G 4

Physically Intrinsic Faithfulness

☆☆☆☆☆

Global and Multi-Modal Interactiveness

☆☆☆☆☆

Stochasticity Planning

☆☆☆☆

World Model

Everywhere

Earth, Jupiter, Venus, Mars...

Ocean, River, Lake...

Seattle, Beijing, Singapore...

Hills, Mountains...

Future World Model

Real-Time

Stochasticity

Zero-Shot

Multi-Scale Planning

Physically Intrinsic Faithfuness

Creative and artistic product

Robotics and agent motion

Autonomous vehicles

Human behaviors

Everything

Ziwei Liu

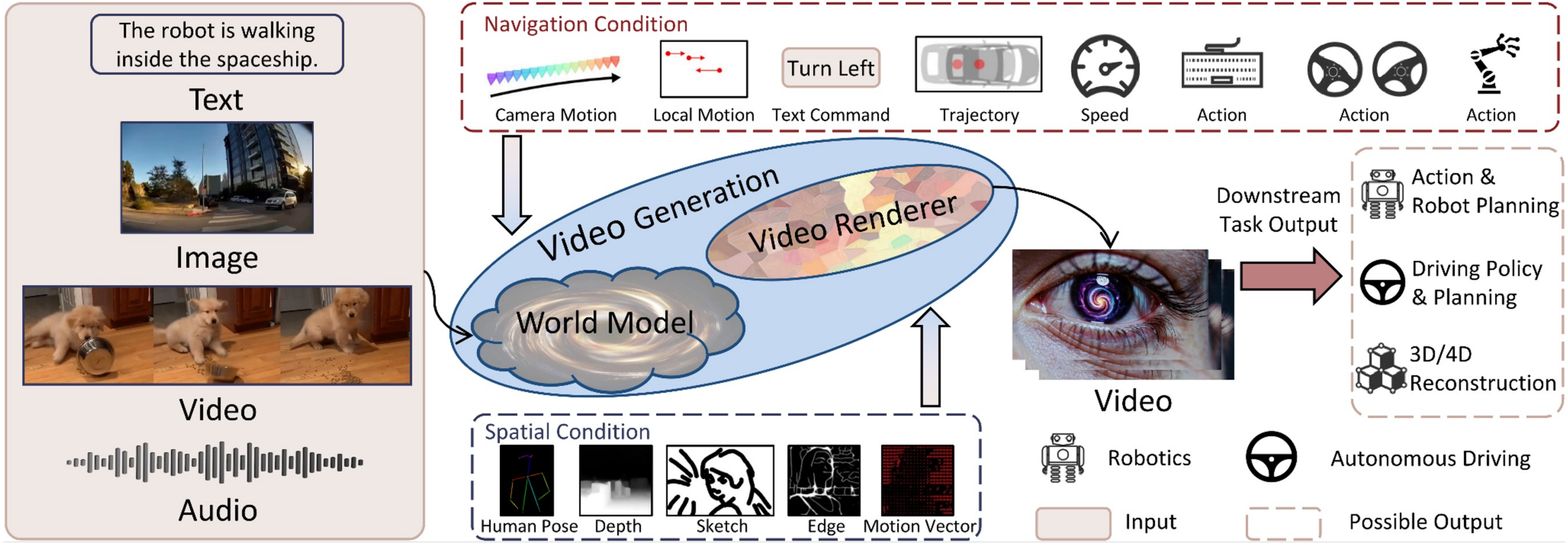
Nanyang Technological University

What is a World Model ?

Discussion Prompt👉 Which is harder: predicting dynamics or generating visuals?

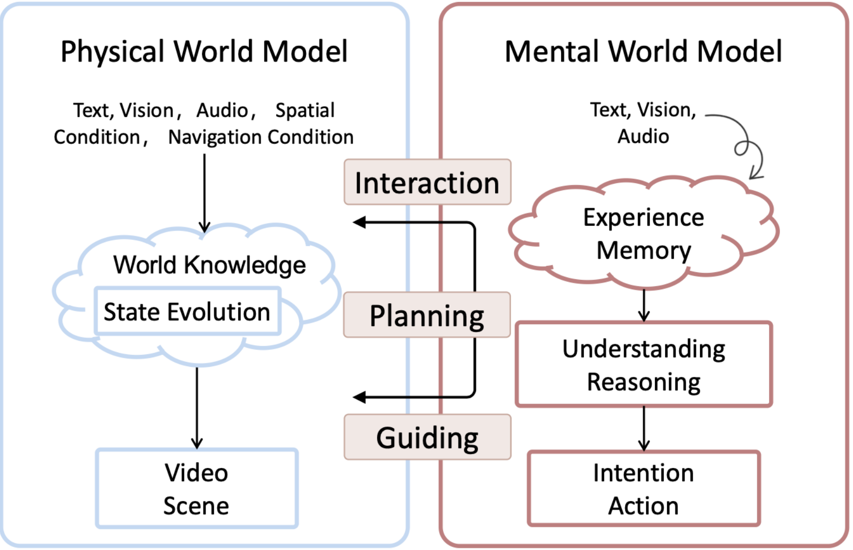
Definition: A **World Model** is a digital engine that encodes knowledge of the environment and simulates its dynamics, with two key components.

1. **Predictor** – learns physical laws & future states
2. **Generator (Renderer)** – renders states into realistic video



Motivation

(Physical & Mental World Model ?)



Prediction & Planning in World Model

- Video generation: good visual fidelity but lacks physics and reasoning
- Real-world needs → predict the future, long-horizon planning
- Vision = key for agents to see, predict, and act (<u>language-centric to vision-centric</u>)
- True world models = simulate dynamics (<u>3D & 4D</u>) + support decisions (<u>digital & physical</u>)

From Generation to Prediction

From Generation → Prediction

- **Generator**: ensures visual consistency
- **Predictor**: models temporal dynamics
- **Prediction Task**: anticipate *next world state*

Takeaway

👉 *Prediction = bridge between perception and causal simulation*

Formalization

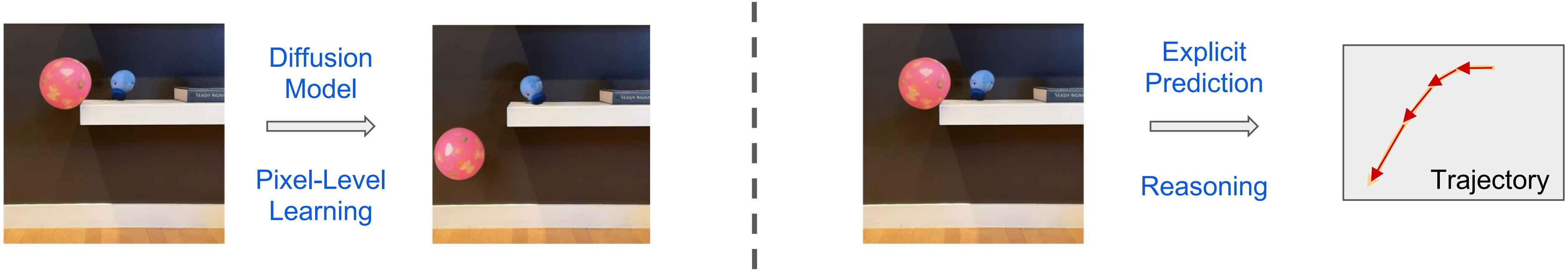
Initial state: $S_0=E(I)$

State transition: $S_{t+1}=F(S_t, I)$

Frame rendering: $V_{t+1}=R(S_{t+1})$

Discussion:

👉 *What is the right information flow? Implicit modeling or explicit modeling?*



Diffusion Model

Pixel-Level Learning

Explicit Prediction

Reasoning

Trajectory

From Generation to Prediction

From Generation → Prediction

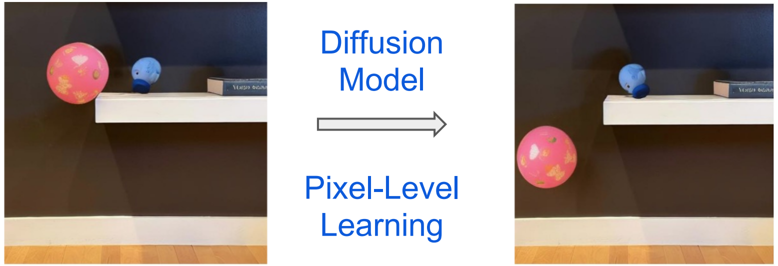
- **Generator**: ensures visual consistency
- **Predictor**: models temporal dynamics
- **Prediction Task**: anticipate *next world state*

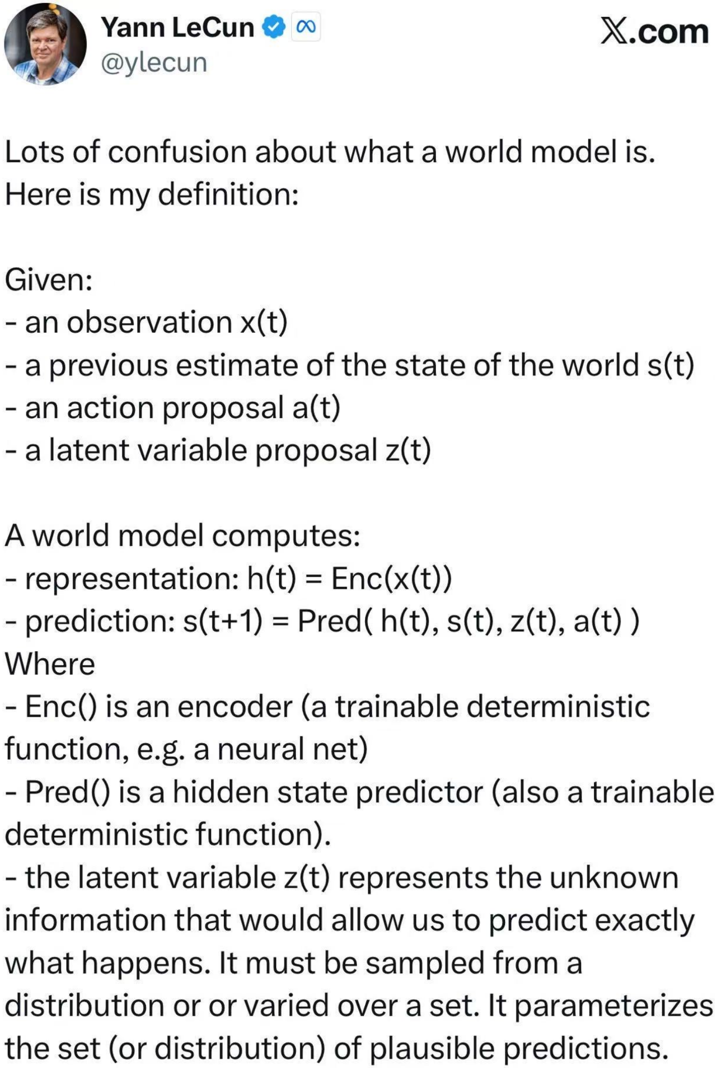
Takeaway

👉 *Prediction = bridge between perception and causal simulation*

Discussion:

👉 *What is the right information flow? Implicit modeling or explicit modeling?*





Yann LeCun
@ylecun

X.com

Lots of confusion about what a world model is.
Here is my definition:

Given:
- an observation x(t)
- a previous estimate of the state of the world s(t)
- an action proposal a(t)
- a latent variable proposal z(t)

A world model computes:
- representation: h(t) = Enc(x(t))
- prediction: s(t+1) = Pred(h(t), s(t), z(t), a(t))
Where
- Enc() is an encoder (a trainable deterministic function, e.g. a neural net)
- Pred() is a hidden state predictor (also a trainable deterministic function).
- the latent variable z(t) represents the unknown information that would allow us to predict exactly what happens. It must be sampled from a distribution or or varied over a set. It parameterizes the set (or distribution) of plausible predictions.

Architectures for Generator + Predictor (Unified Model ?)

Typical Architectures

Component	Methods	Examples
Generator 🎥	Diffusion Models / Autoregressive	High-Quality Videos
Predictor ⚙️	Latent Transition / Physics-Informed	3D State Dynamics
Joint Models 🔗	UniSim, Drive-WM, Cosmos, Genie	Unified World Models

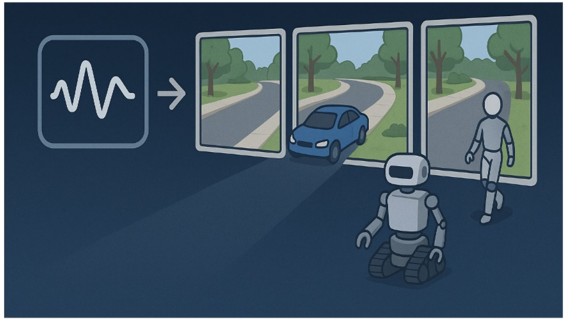
Key Insight:

👉 *Generator + Predictor must be tightly coupled*

Discussion：

👉 *Should we build them separately or as one unified model?*

How Planning Emerges (CoT ?)



From Prediction to Planning

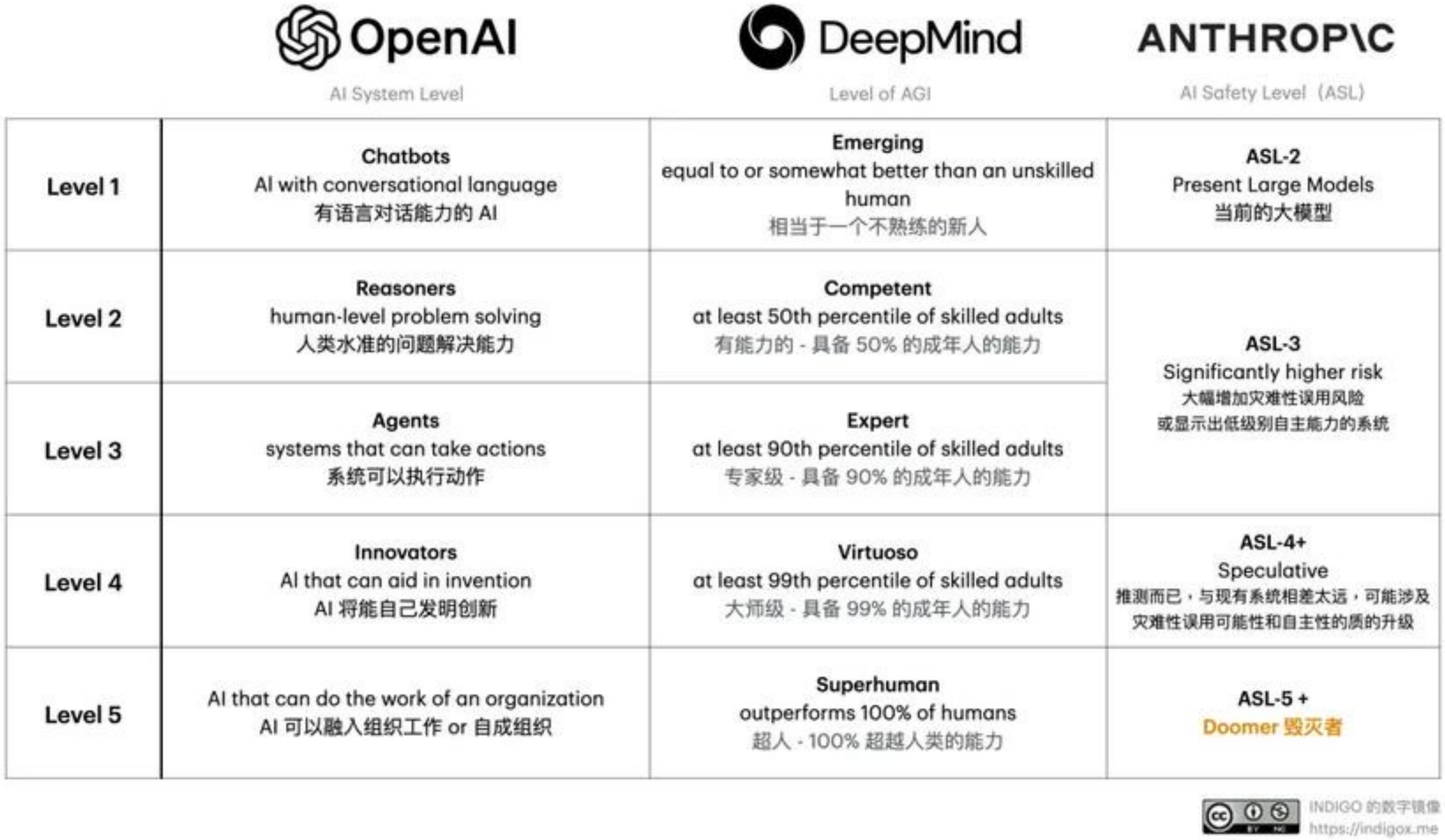
- **Prediction**: Anticipate the *next state*
  Planning: Chain of Predictions to achieve *goals*
- **Core Idea**: Evaluate Alternative Futures ⇨ Conduct Best Action

Applications:

🚗 **Autonomous Driving**: Predict trajectories of cars & pedestrians, then plan safe driving.
🤖 **Robotics**: Plan navigation / action steps.
🎮 **Gaming / Agents**: Predict opponent moves and plan strategies accordingly.

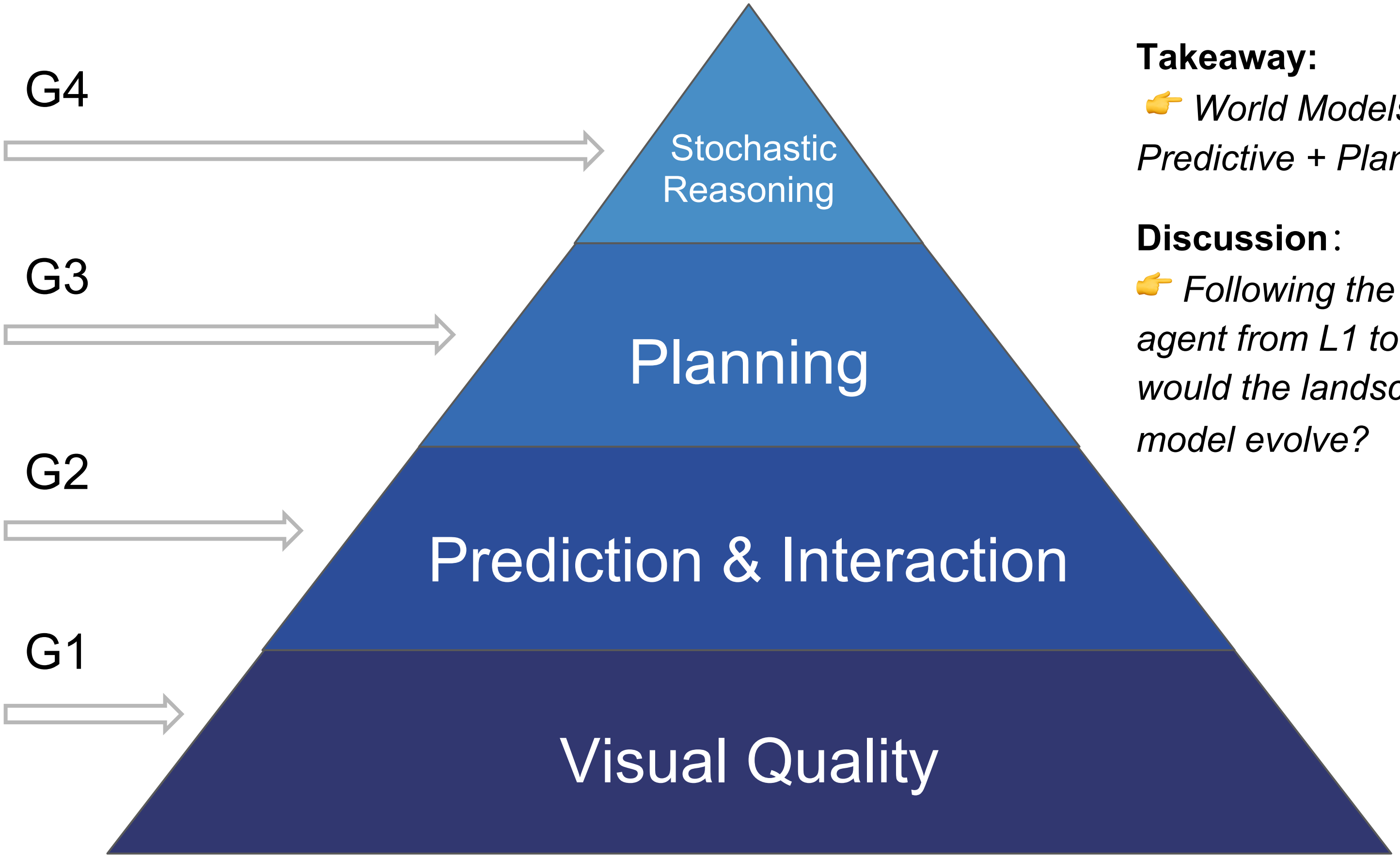
Discussion：
👉 *What is the bottleneck of “next token prediction” moment for prediction? Data, architecture or objective?*

The Five Levels of AGI



	OpenAI AI System Level	DeepMind Level of AGI	ANTHROP\C AI Safety Level（ASL）
Level 1	**Chatbots** AI with conversational language 有语言对话能力的 AI	**Emerging** equal to or somewhat better than an unskilled human 相当于一个不熟练的新人	**ASL-2** Present Large Models 当前的大模型
Level 2	**Reasoners** human-level problem solving 人类水准的问题解决能力	**Competent** at least 50th percentile of skilled adults 有能力的 - 具备 50% 的成年人的能力	**ASL-3** Significantly higher risk 大幅增加灾难性误用风险 或显示出低级别自主能力的系统
Level 3	**Agents** systems that can take actions 系统可以执行动作	**Expert** at least 90th percentile of skilled adults 专家级 - 具备 90% 的成年人的能力	
Level 4	**Innovators** AI that can aid in invention AI 将能自己发明创新	**Virtuoso** at least 99th percentile of skilled adults 大师级 - 具备 99% 的成年人的能力	**ASL-4+** Speculative 推测而已，与现有系统相差太远，可能涉及灾难性误用可能性和自主性的质的升级
Level 5	AI that can do the work of an organization AI 可以融入组织工作 or 自成组织	**Superhuman** outperforms 100% of humans 超人 - 100% 超越人类的能力	**ASL-5 +** Doomer 毁灭者

The Four Generations of World Model

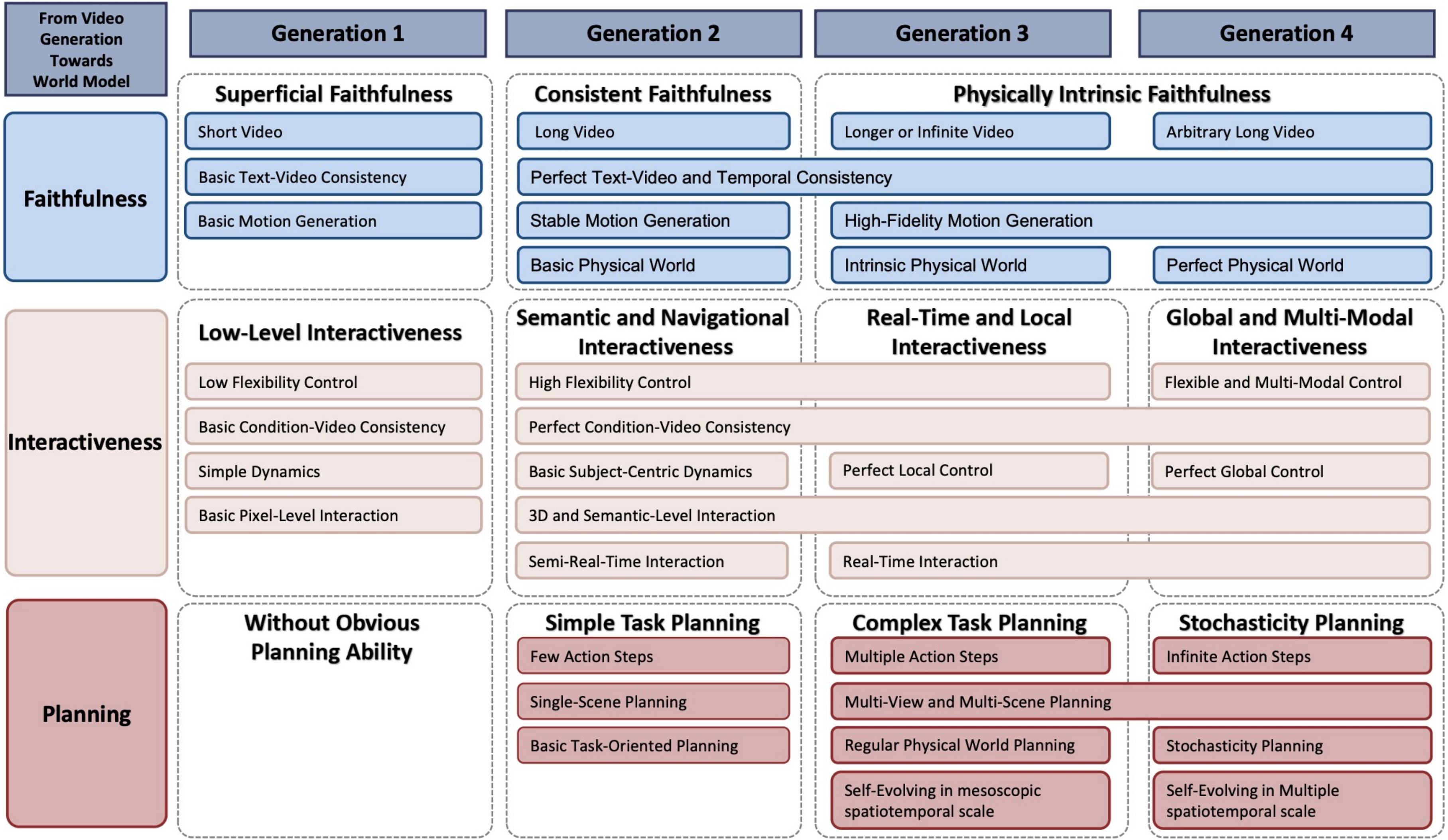


Takeaway:

👉 *World Models : Generative + Predictive + Planning*

Discussion:

👉 *Following the progress of AI agent from L1 to L4, how fast would the landscape of world model evolve?*



From Video Generation Towards World Model	Generation 1	Generation 2	Generation 3	Generation 4
Faithfulness	Superficial Faithfulness	Consistent Faithfulness	Physically Intrinsic Faithfulness	
	Short Video	Long Video	Longer or Infinite Video	Arbitrary Long Video
	Basic Text-Video Consistency	Perfect Text-Video and Temporal Consistency		
	Basic Motion Generation	Stable Motion Generation	High-Fidelity Motion Generation	
		Basic Physical World	Intrinsic Physical World	Perfect Physical World
Interactiveness	Low-Level Interactiveness	Semantic and Navigational Interactiveness	Real-Time and Local Interactiveness	Global and Multi-Modal Interactiveness
	Low Flexibility Control	High Flexibility Control		Flexible and Multi-Modal Control
	Basic Condition-Video Consistency	Perfect Condition-Video Consistency		
	Simple Dynamics	Basic Subject-Centric Dynamics	Perfect Local Control	Perfect Global Control
	Basic Pixel-Level Interaction	3D and Semantic-Level Interaction		
		Semi-Real-Time Interaction	Real-Time Interaction	
Planning	Without Obvious Planning Ability	Simple Task Planning	Complex Task Planning	Stochasticity Planning
		Few Action Steps	Multiple Action Steps	Infinite Action Steps
		Single-Scene Planning	Multi-View and Multi-Scene Planning	
		Basic Task-Oriented Planning	Regular Physical World Planning	Stochasticity Planning
			Self-Evolving in mesoscopic spatiotemporal scale	Self-Evolving in Multiple spatiotemporal scale

Open Challenges

🔗

Coupling of Generator & Predictor

Unified Model Needed?

⏳

Long-Horizon Consistency vs. Efficiency

Accuracy vs Computational Resources

🎲

Stochasticity-Aware Planning

Handle Multiple Futures

🎯

Evaluation & Benchmarks

Need Better Metrics

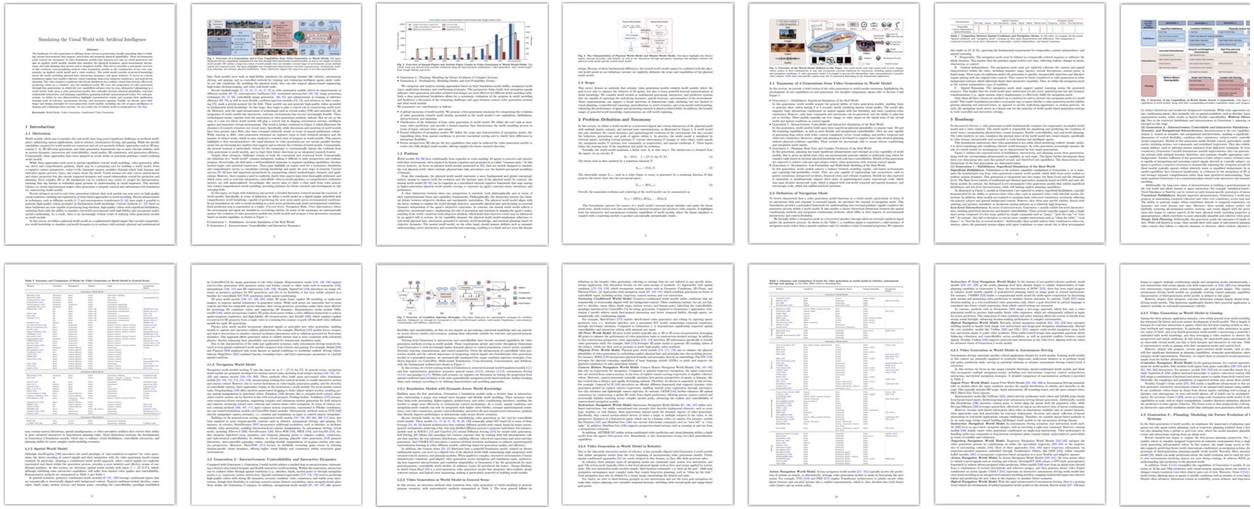
Conclusion & Call for Discussion

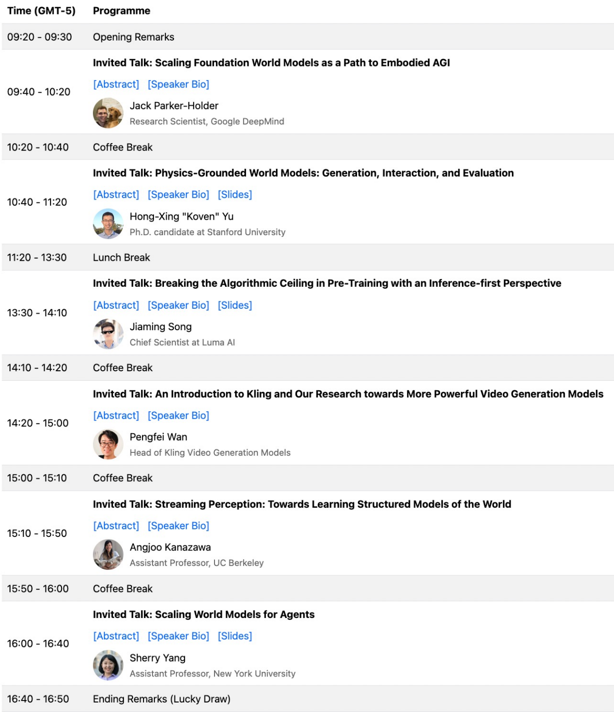
- The field of world models is evolving from **generation** → **prediction** → **planning**.

- **Integration of prediction and planning** is key to building more robust and intelligent systems.

- We look forward to **further discussion and insights** in Mini3DV.

More Resources: https://world-model-tutorial.github.io/





Time (GMT-5)	Programme
09:20 - 09:30	Opening Remarks
09:40 - 10:20	**Invited Talk: Scaling Foundation World Models as a Path to Embodied AGI** [Abstract] [Speaker Bio] Jack Parker-Holder Research Scientist, Google DeepMind
10:20 - 10:40	Coffee Break
10:40 - 11:20	**Invited Talk: Physics-Grounded World Models: Generation, Interaction, and Evaluation** [Abstract] [Speaker Bio] [Slides] Hong-Xing "Koven" Yu Ph.D. candidate at Stanford University
11:20 - 13:30	Lunch Break
13:30 - 14:10	**Invited Talk: Breaking the Algorithmic Ceiling in Pre-Training with an Inference-first Perspective** [Abstract] [Speaker Bio] [Slides] Jiaming Song Chief Scientist at Luma AI
14:10 - 14:20	Coffee Break
14:20 - 15:00	**Invited Talk: An Introduction to Kling and Our Research towards More Powerful Video Generation Models** [Abstract] [Speaker Bio] Pengfei Wan Head of Kling Video Generation Models
15:00 - 15:10	Coffee Break
15:10 - 15:50	**Invited Talk: Streaming Perception: Towards Learning Structured Models of the World** [Abstract] [Speaker Bio] Angjoo Kanazawa Assistant Professor, UC Berkeley
15:50 - 16:00	Coffee Break
16:00 - 16:40	**Invited Talk: Scaling World Models for Agents** [Abstract] [Speaker Bio] [Slides] Sherry Yang Assistant Professor, New York University
16:40 - 16:50	Ending Remarks (Lucky Draw)