

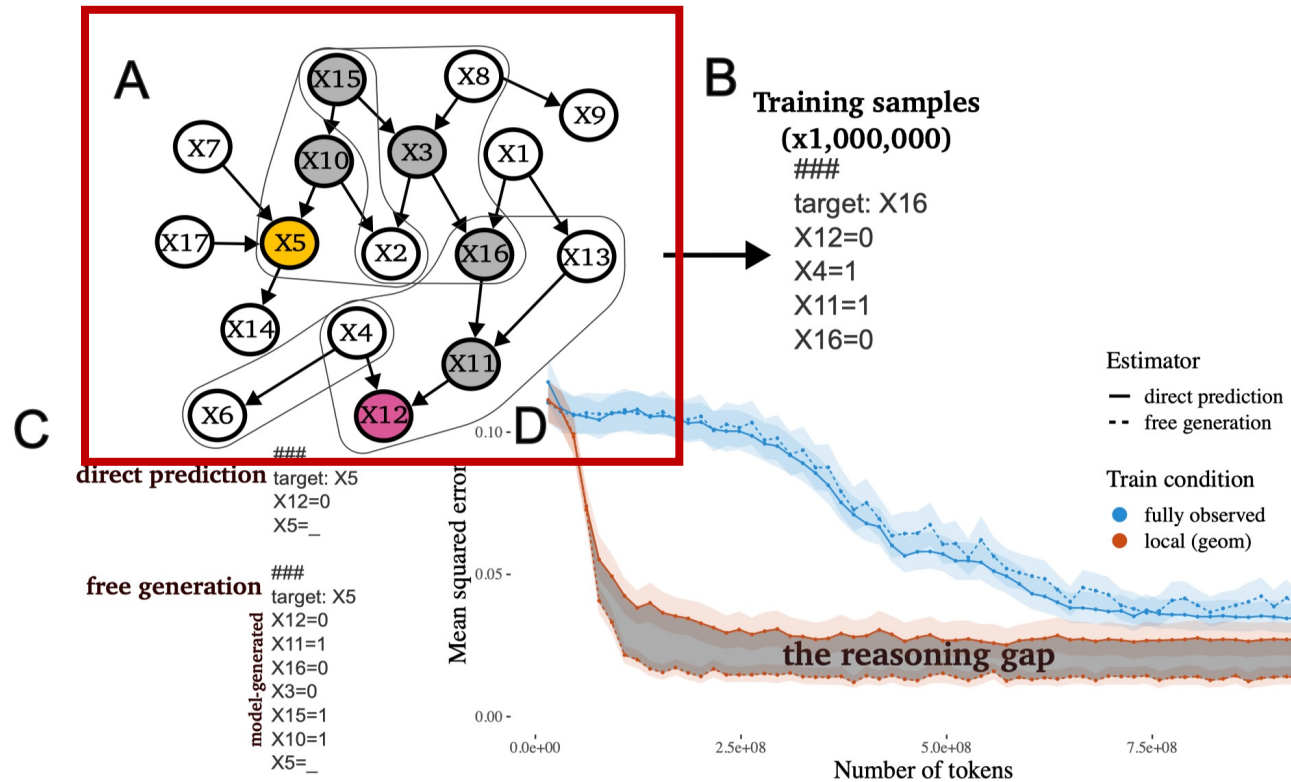
Bridging World Models and Multimodal Reasoning

Jialong Wu

School of Software, Tsinghua University

June 4th, 2026

Reasoning = World Model + Search Strategy

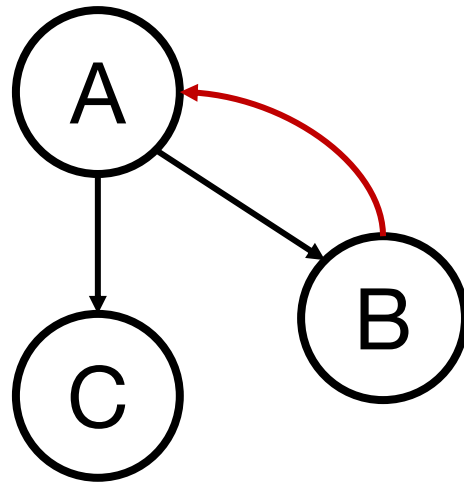


Chain-of-Thought =
Flattened searching trajectory

Why think step by step? Reasoning emerges from the locality of experience. NeurIPS 2023.

Reasoning = World Model + Search Strategy

Mathematical reasoning:
Searching on **a model in the mind**



If A then B, ...
Wait !

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a + x}})^2 = x^2 \Rightarrow a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \Rightarrow a^2 - 2ax^2 + (x^2)^2 = a + x \Rightarrow x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step by step to identify whether the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \Rightarrow \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

Reasoning = World Model + Search Strategy

Planning:

Searching on a model
of external world

Training Example | TravelPlanner

Task: Please help me plan a trip from St. Petersburg to Rockford spanning 3 days from March 16th to March 18th, 2022. The travel should be planned for a single person with a budget of \$1,700.

Current State: Total Days: 3, Initial Budget: \$1700, Spent: \$0, Remaining: \$1700 Day 1: transportation: PENDING, breakfast: PENDING, attraction: PENDING, lunch: PENDING, dinner: PENDING, accommodation: PENDING Next action required: Plan day 1 transportation Available: Flight F3573659: \$474, Departure 15:40, Arrival 17:04

Expert Action: {"action_type": "SET_TRANSPORTATION", "day": 1, "field": "transportation", "value": "Flight Number: F3573659, from St. Petersburg to Rockford, Departure Time: 15:40, Arrival Time: 17:04", "cost": 474}

Implicit World Modeling:

- SET_TRANSPORTATION (F3573659) → “After this action, you’ve spent \$474, leaving \$1226 from your \$1700 budget (28% used). Good start! You’re traveling to your destination.”
- SKIP_TRANSPORTATION → “After this action, you’ve spent \$0, leaving \$1700 from your \$1700 budget (0% used).”

Self-Reflection: Important: Consider all constraints. Flight F3573659 is a valid option at \$474. It satisfies all constraints and gets me to my destination efficiently. The only alternative is skipping transportation, which prevents reaching the destination. Given all constraints and optimization factors, Flight F3573659 is indeed the best choice as it’s the most cost-effective option that enables the trip. **Therefore, Flight F3573659 is optimal.**

Agent Learning via Early Experience.

Contents

Bridging:

- **World Models**
- Multimodal Reasoning

Incomplete History of World Models

World Models

A Path Towards Autonomous Machine Intelligence

Version 0.9.2, 2022-06-27

David Ha¹ Jürgen Schmidhuber^{2,3}



?

→ 2018

→ 2019

→ 2022

→ 2024

→ 2026



DREAM TO CONTROL: LEARNING BEHAVIORS
BY LATENT IMAGINATION



Dudes showing up at the first
conference for World Models...



Mental Models of Human



Kenneth Craik
Psychologist
(1943)

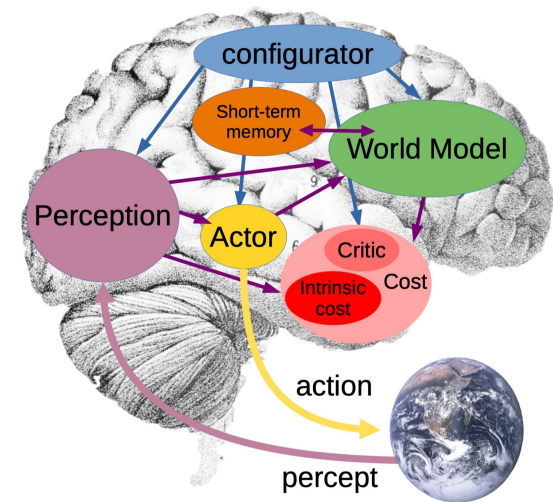
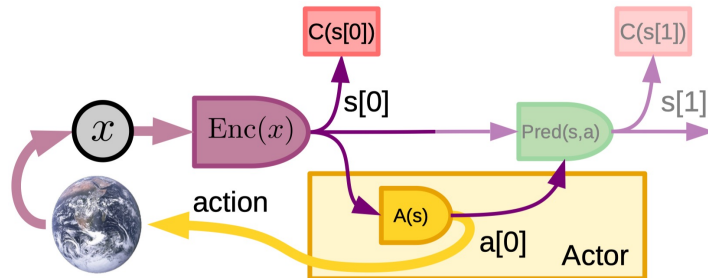
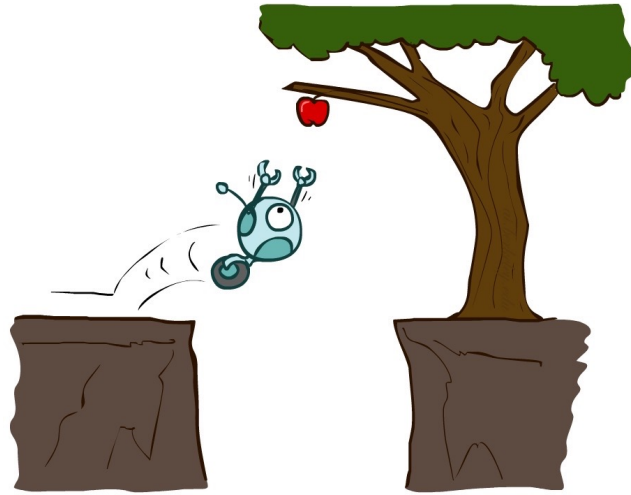
Simulation and Planning:

“ If the organism carries a “small-scale model” of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way react in a much fuller, safer, and more competent manner to the emergencies which face it.”

World Models for AI: From System-1 to System-2

System-1 Agent (Reflex):

Not utilize
the world model
nor the cost.



World Models:

How the world transits under interventions

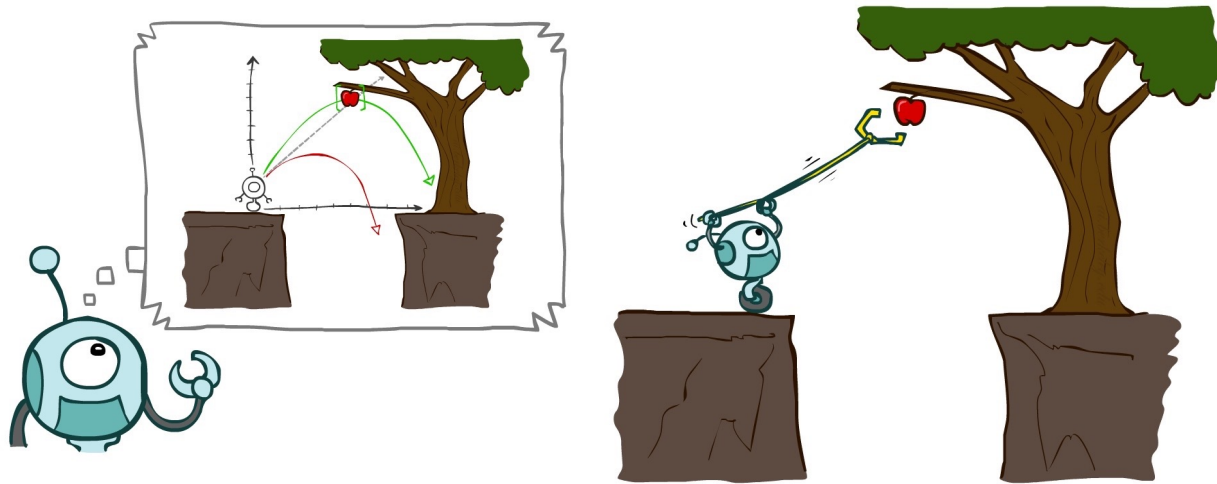
$f: \text{State} \times \text{Action} \rightarrow \text{Next State}$

Interventions

Yann LeCun. A path towards autonomous machine intelligence. 2022.

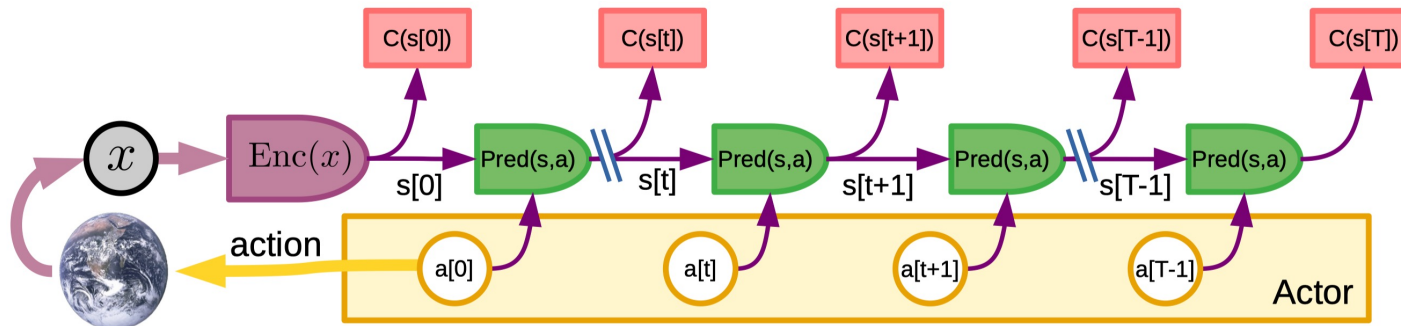
Dan Klein and Pieter Abbeel. Introduction to Artificial Intelligence.

World Models for AI: From System-1 to System-2



System-2 Agent (Planning/Model Predictive Control):

Act through an optimization procedure running the **world model**.



Amortized Inference (Model-based RL):

A **policy** module mimicking (distilling) the planned optimal actions

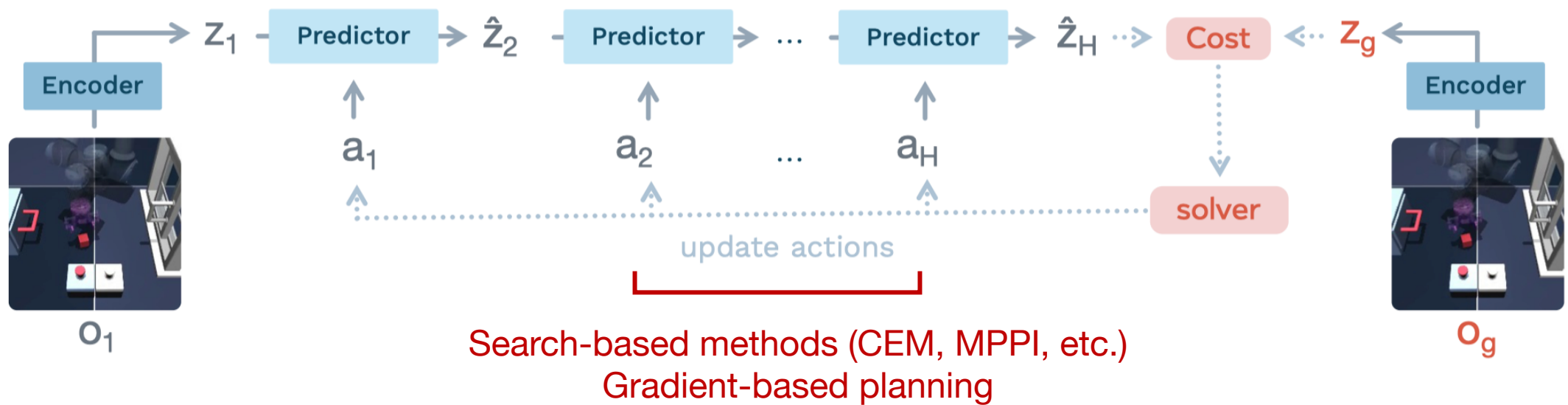
Yann LeCun. A path towards autonomous machine intelligence. 2022.

Dan Klein and Pieter Abbeel. Introduction to Artificial Intelligence.

Planning with World Models

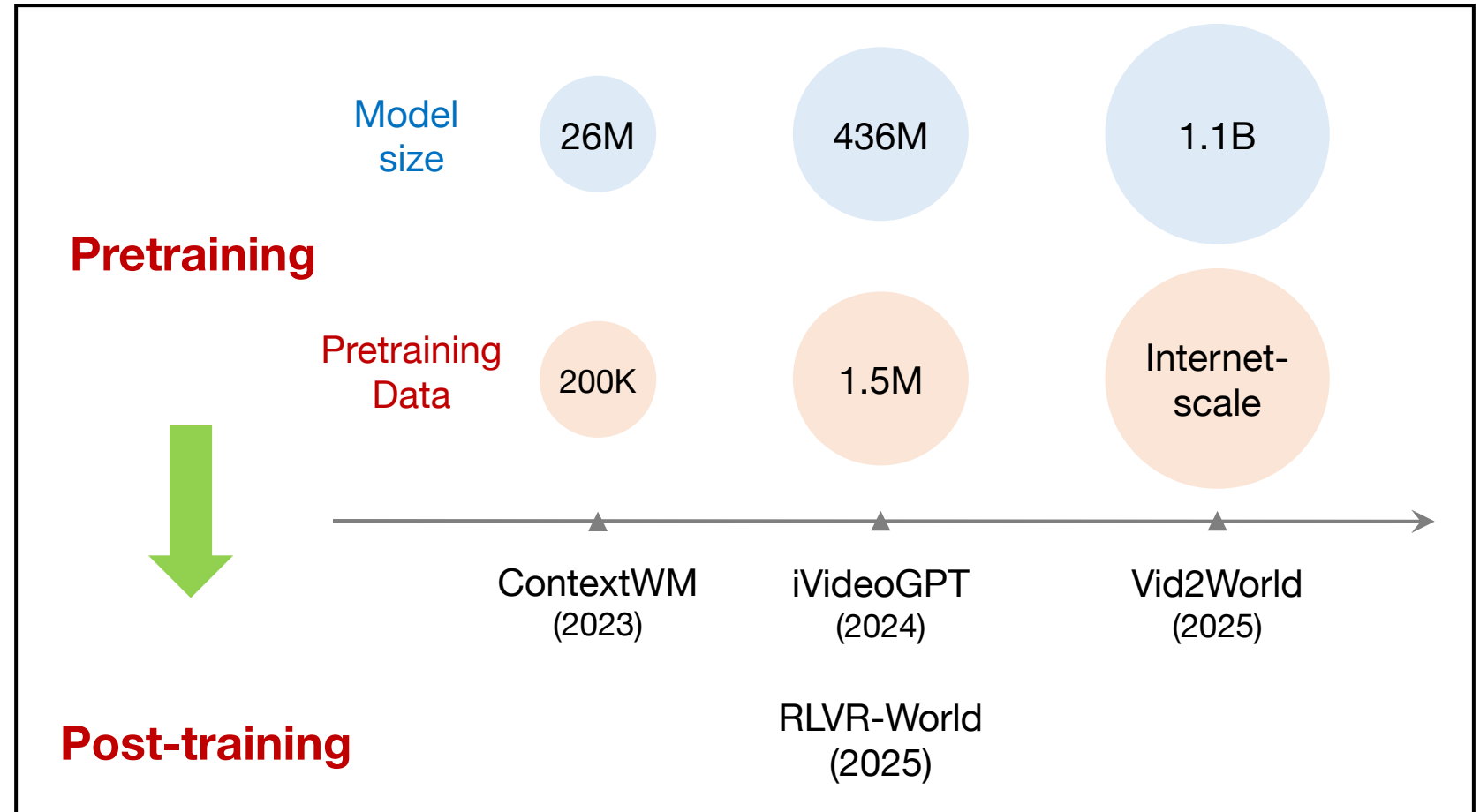
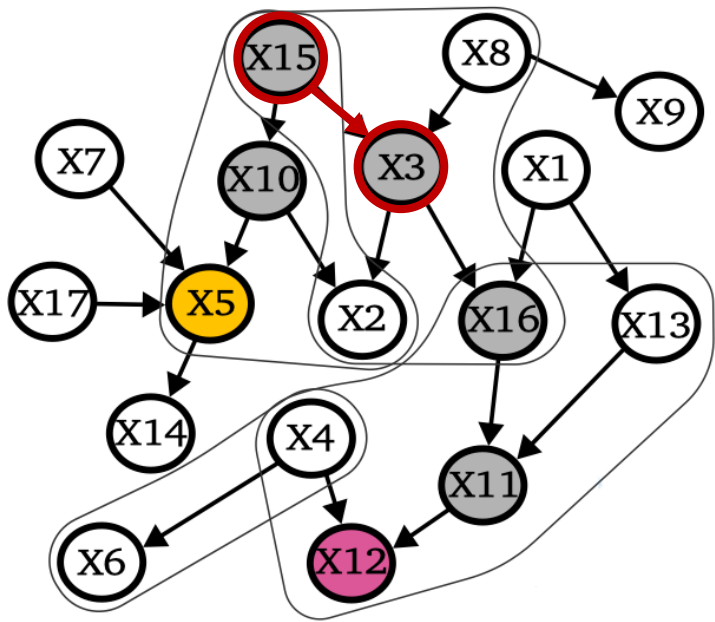
Optimizing Action Sequence: Once you can evaluate it, you can improve it!

$$a^* = \underset{a_{1:T}}{\operatorname{argmax}} \mathbb{E}_{s_{t+1}, r_t \sim p_\theta(s_t, a_t)} \left[\sum_t r_t \right]$$

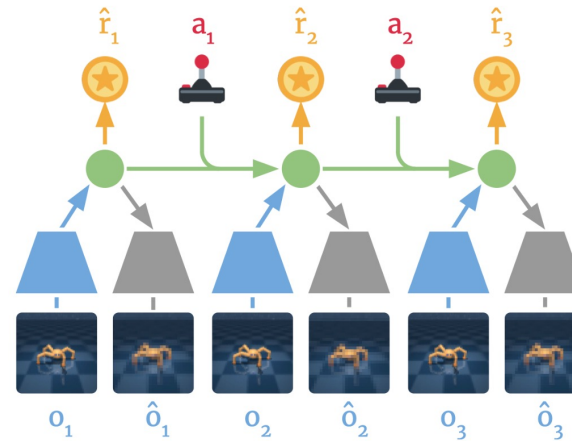
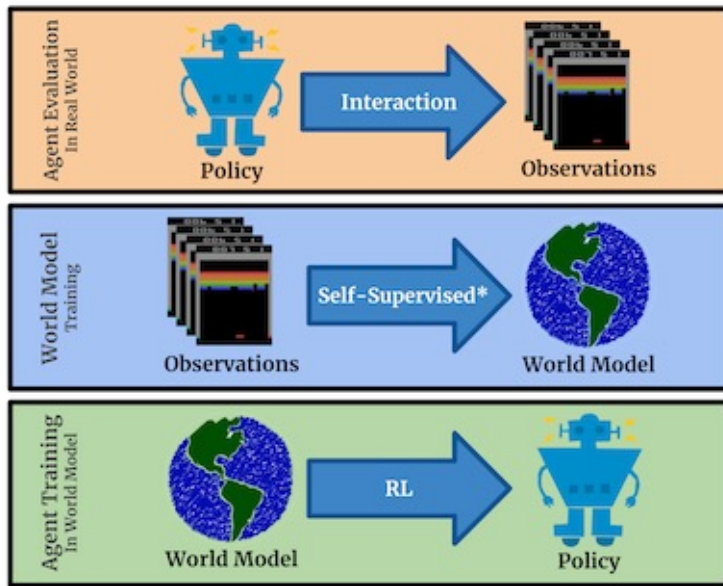


LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels.

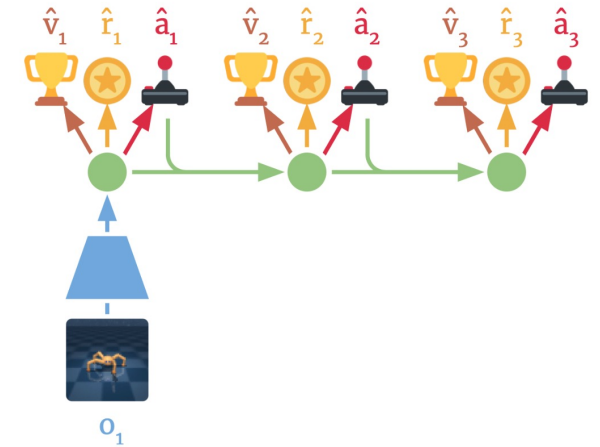
Making World Models More Scalable and High-Fidelity



Dreamer: An Instantiation of World Models



Model Learning with
Sequential Variational Inference



Behavior Learning by
Latent Imagination

Lukasz Kaiser, et al. Model-Based Reinforcement Learning for Atari. ICLR 2020.

Danijar Hafner, et al. Dream to control: Learning behaviors by latent imagination. ICLR 2020.

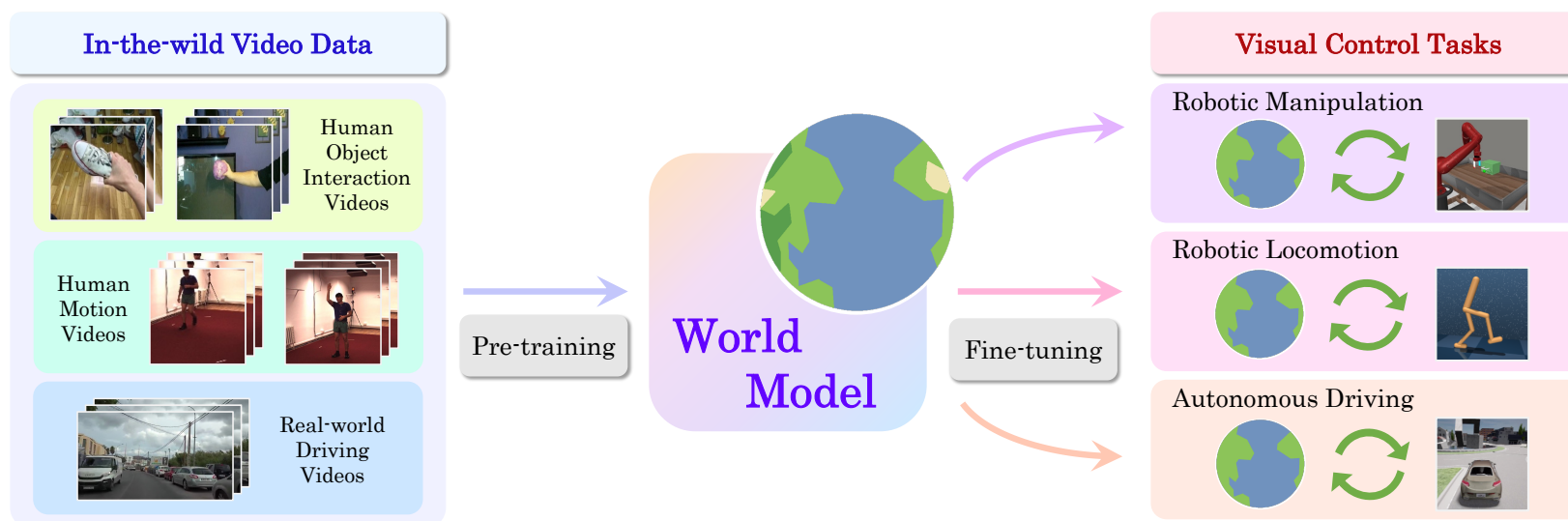
Danijar Hafner, et al. Mastering atari with discrete world models. ICLR 2021.

Danijar Hafner, et al. Mastering diverse control tasks through world models. Nature 2025.

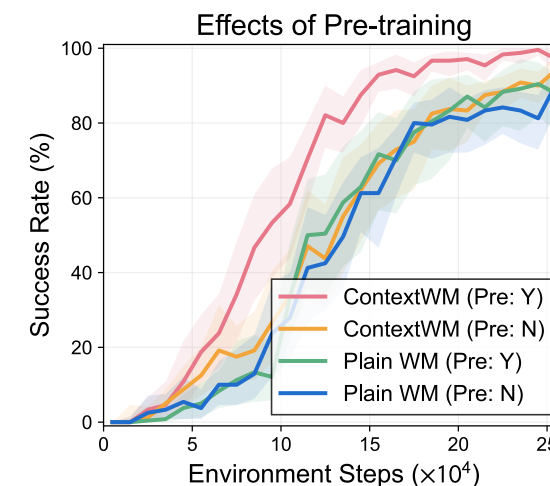
Kick Start Video Pre-training: ContextWM (2023)

Video Pre-training for World Models:

Learning **representations** useful for understanding the **dynamics** via **generative pretraining** on **abundant in-the-wild videos**

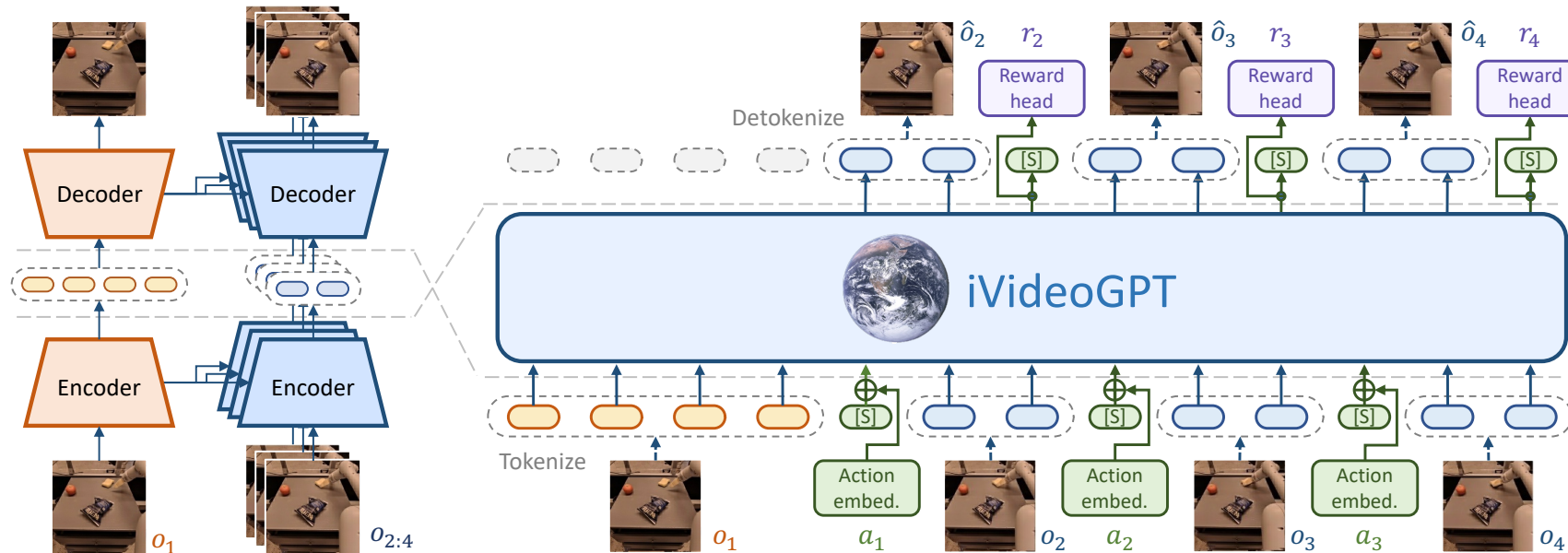


Positive transfer for MBRL sample efficiency



Scalable & Interactive World Models: iVideoGPT (2024)

ContextWM based on recurrent networks (**RNNs**) still **lack the scalability** required by large-scale data...



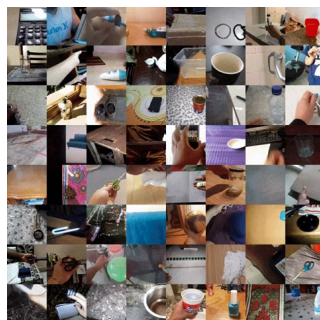
✓ Scalability
✓ Interactivity

Compressive tokenization

Interactive prediction with Transformers

Scalable & Interactive World Models: iVideoGPT (2024)

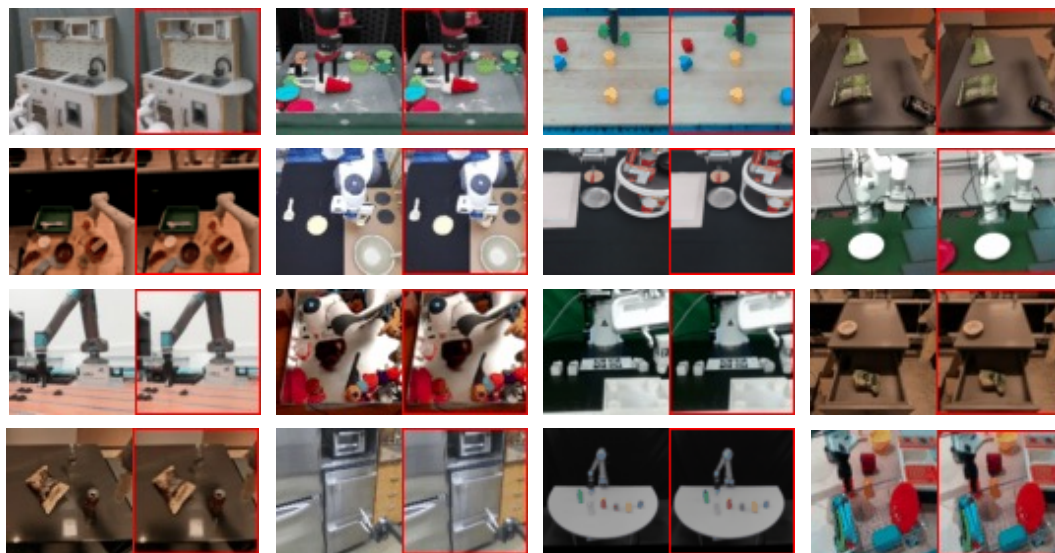
Total 1.5 million trajectories of human and robot manipulation



A versatile foundation adaptable for a wide range of downstream tasks

Pre-training

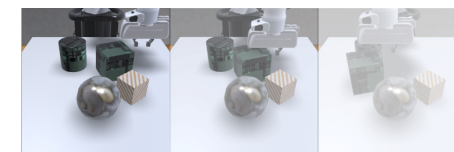
Action-free video prediction samples



Video Prediction



Visual Planning



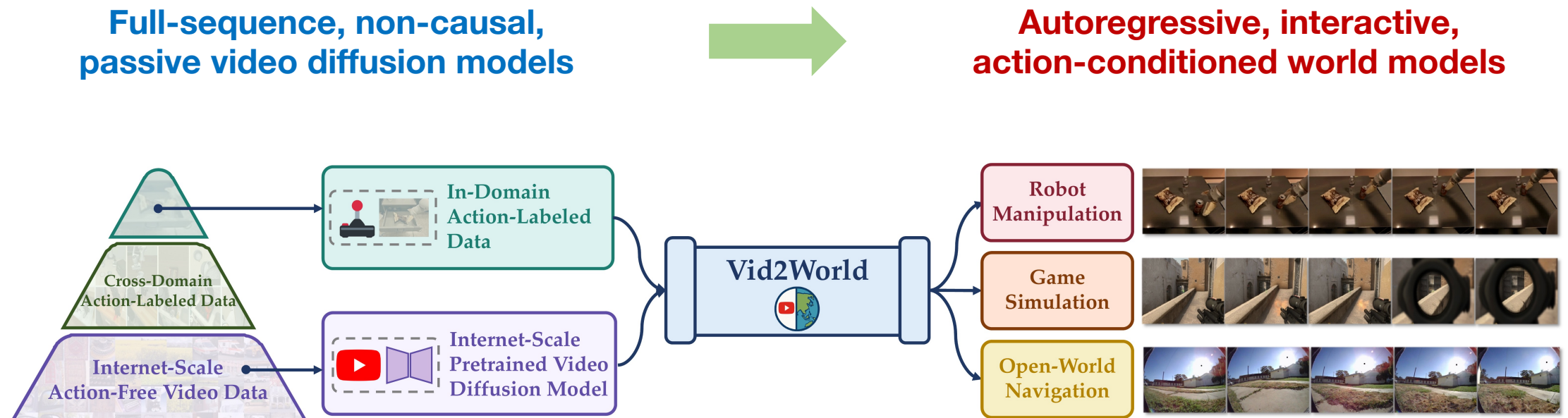
Visual Model-based RL



Leveraging Extensive Pre-training: Vid2World (2025)

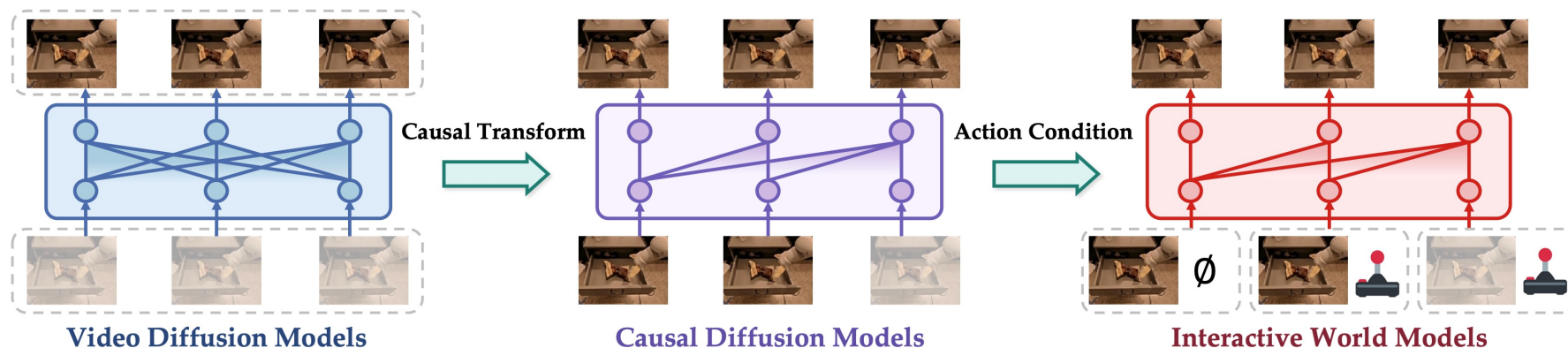
We are still **limited by domain-specific pre-training...**

Vid2World transfers the physical priors and generative capabilities learned by **Internet-scale pre-trained** video diffusion models.



Leveraging Extensive Pre-training: Vid2World (2025)

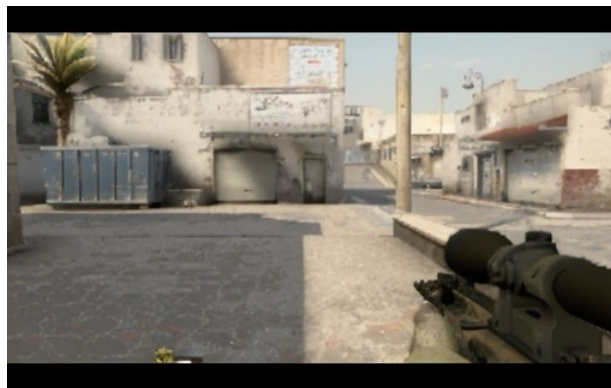
- ✓ Video Diffusion
Causalization
- ✓ Causal Action
Guidance



Robot manipulation



3D game simulation



Open-world navigation

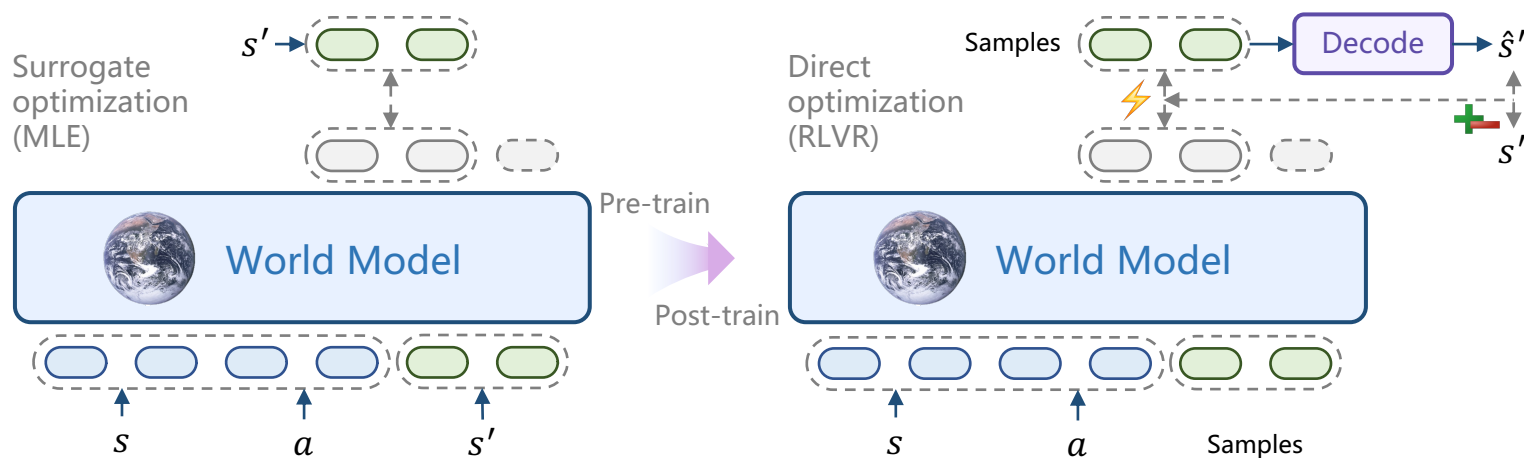


Siqiao Huang, [Jialong Wu](#), et al. Vid2World: Crafting Video Diffusion Models to Interactive World Models. ICLR 2026.

Post-training World Models with RL: **RLVR-World** (2025)

Pre-training: Maximum likelihood estimation (MLE) via cross-entropy loss is **scalable** but often **misalign** with task-specific goals of world models

- Mean squared error vs. perceptual metrics
 - Teacher forcing vs. autoregressive generation
 - Non-end-to-end architecture
- ➔
- Language models: **repetition** and **hallucination**
 - Video models: **blurry** predictions



Post-training: Reinforcement Learning with Verifiable Rewards (RLVR)

$$R_i = \text{sign}(D) \cdot D(\hat{s}'_i, s')$$

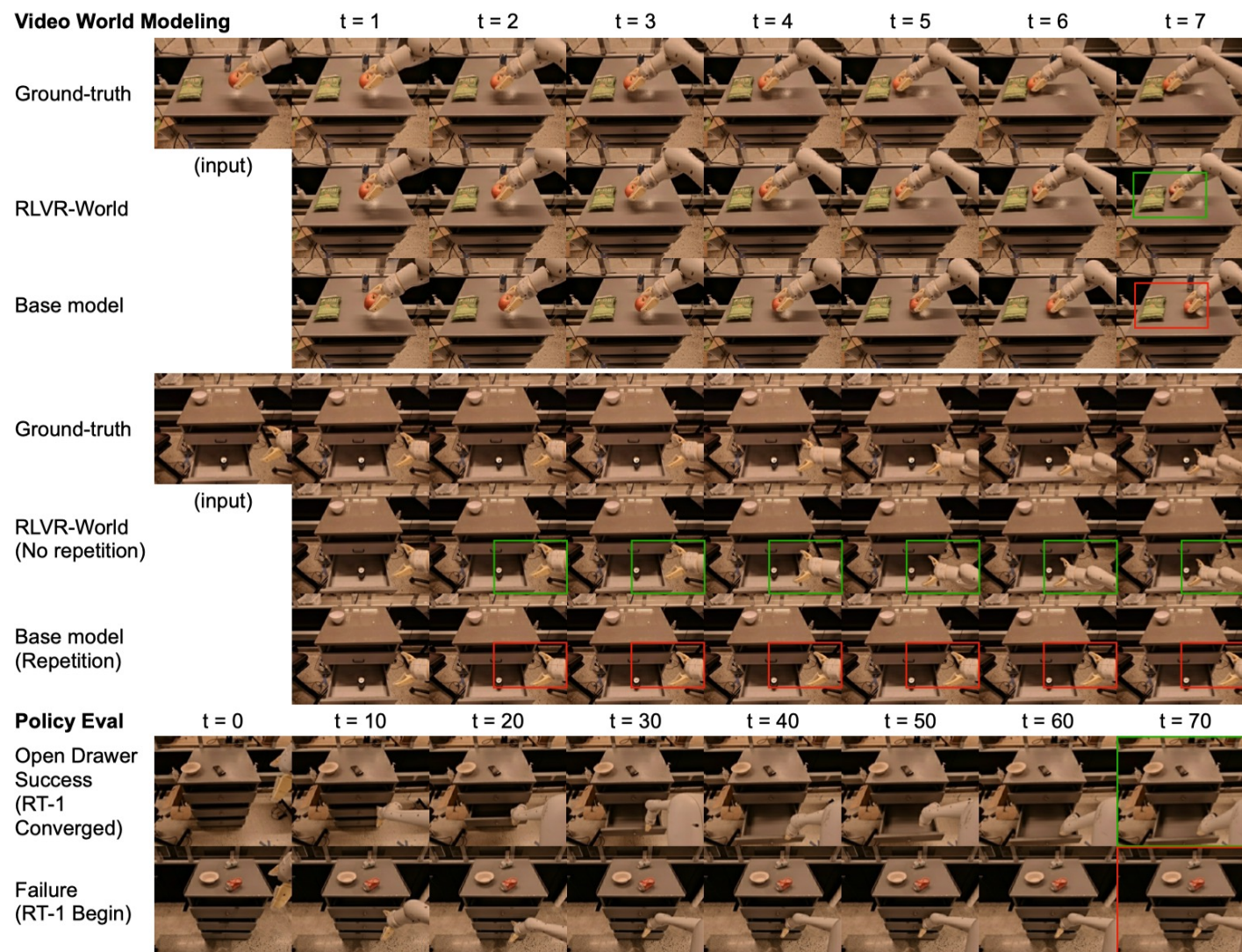
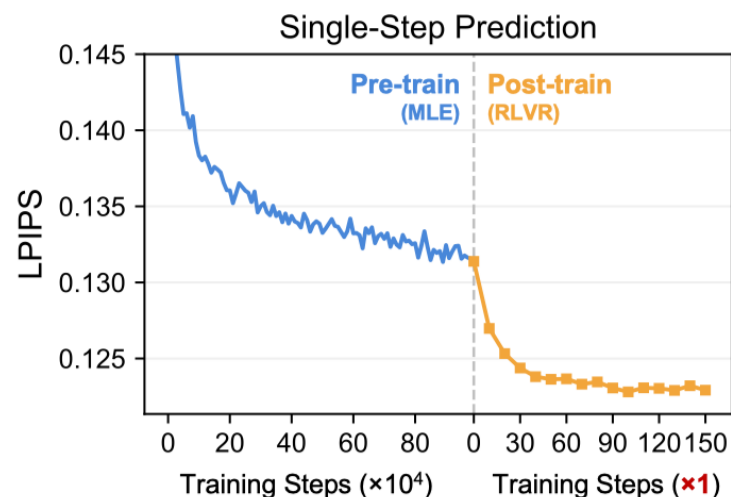
Prediction Metrics as Verifiable Rewards

Post-training World Models with RL: **RLVR-World** (2025)

iVideoGPT with RLVR

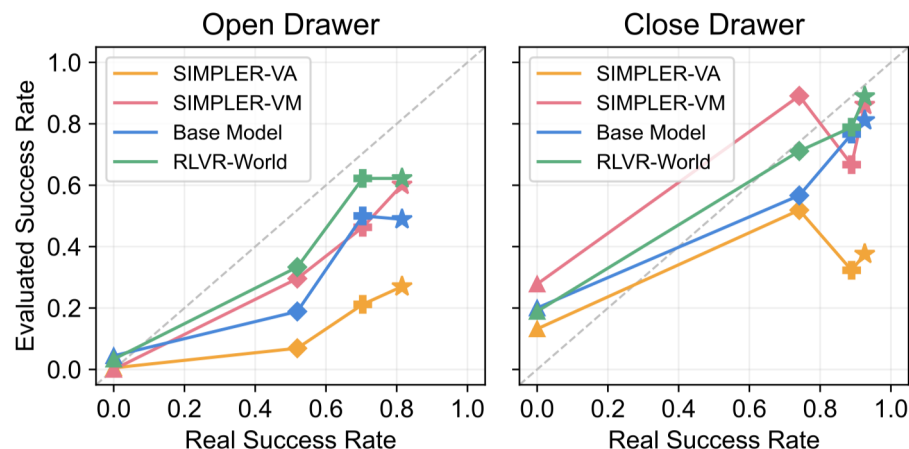
Sharp training speed up:
More than **1000x**!

Reduced Artifacts:
Repetition caused by likelihood-based objectives

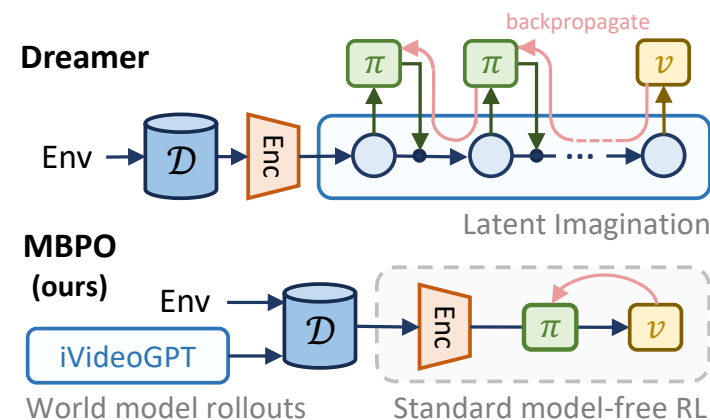


World Model Utilities: Evaluation, Planning and Improvement

Real2Sim Policy Evaluation

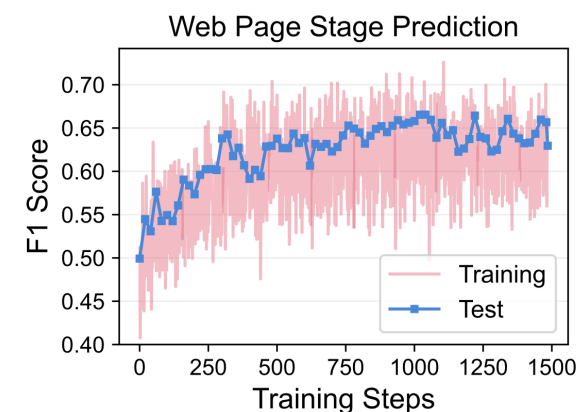


Visual Model-based RL with iVideoGPT



Model Predictive Control for Web Agents

Model	Web Page State Prediction			Web Agent Success Rate
	Precision	Recall	F1	
Base (R1-Dist.-Qwen-1.5B)	15.59%	15.70%	11.83%	n/a
SFT	48.99%	56.05%	49.94%	12.06%
RLVR-World (Ours)	72.77%	64.55%	65.11%	14.29%
Δ	+48.5%	+15.1%	+30.3%	+18.4%

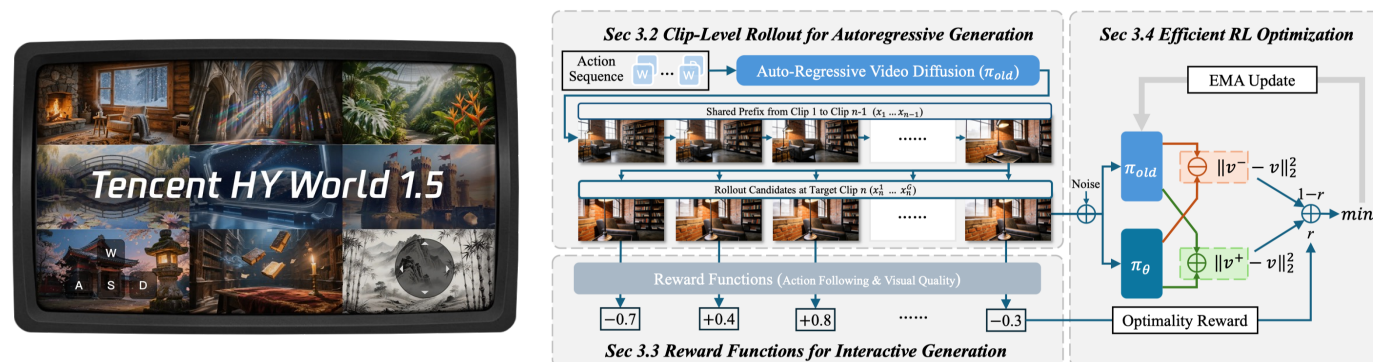


Jialong Wu, et al. iVideoGPT: Interactive VideoGPTs are Scalable World Models. NeurIPS 2024.

Jialong Wu, et al. RLVR-World: Training World Models with Reinforcement Learning. NeurIPS 2025.

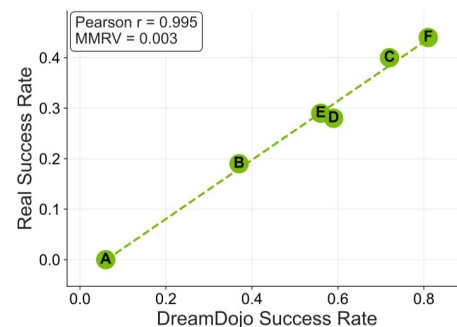
Fast-Growing Field of World Models

Post-training World Models with RL

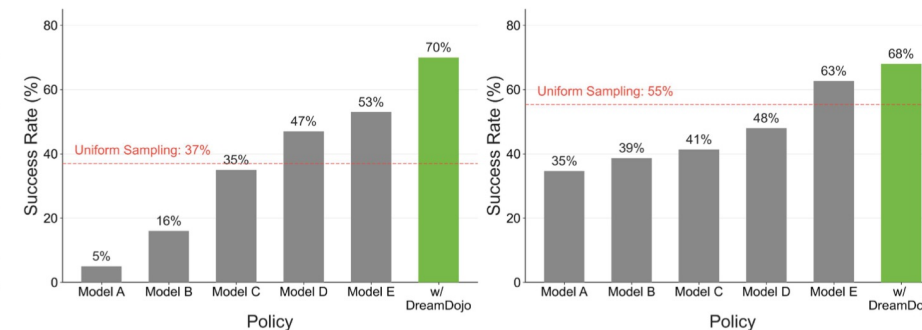


World Model as An Environment for Policy Evaluation and Improvement

(a) Real vs. DREAMDOJO success rates.

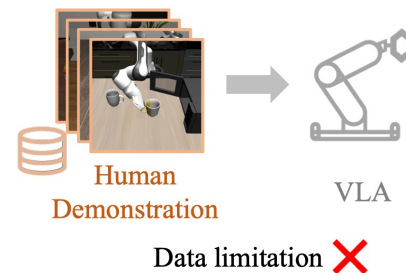


(b) Model-based planning results.

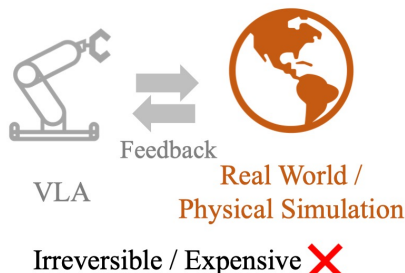


Post-training VLA with World Models

(a) Imitation Learning



(b) Previous Post Training



(c) Post Training with World-Env



Fast-Growing Field of World Models



Are We Already Done?

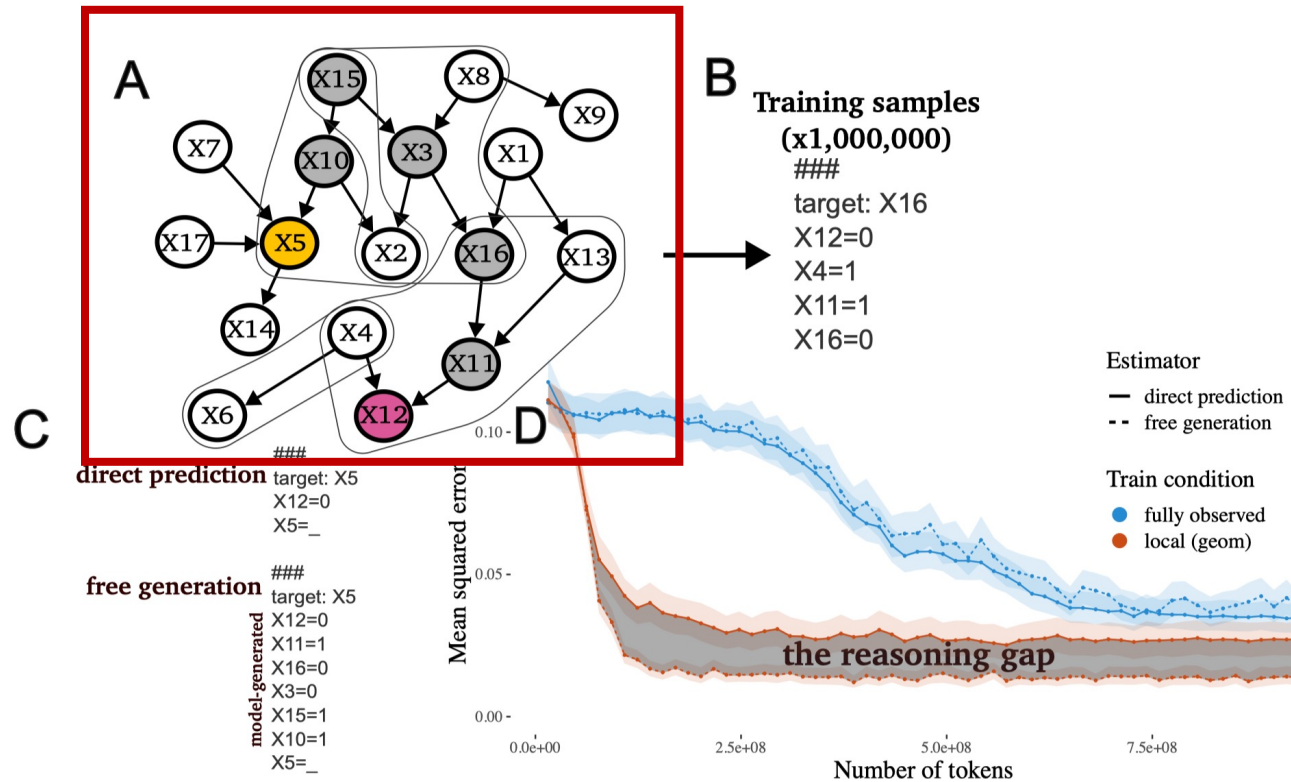
Humans do not think in terms of
continuous raw-pixel prediction or **algorithmic action optimization**

Contents

Bridging:

- World Models
- **Multimodal Reasoning**

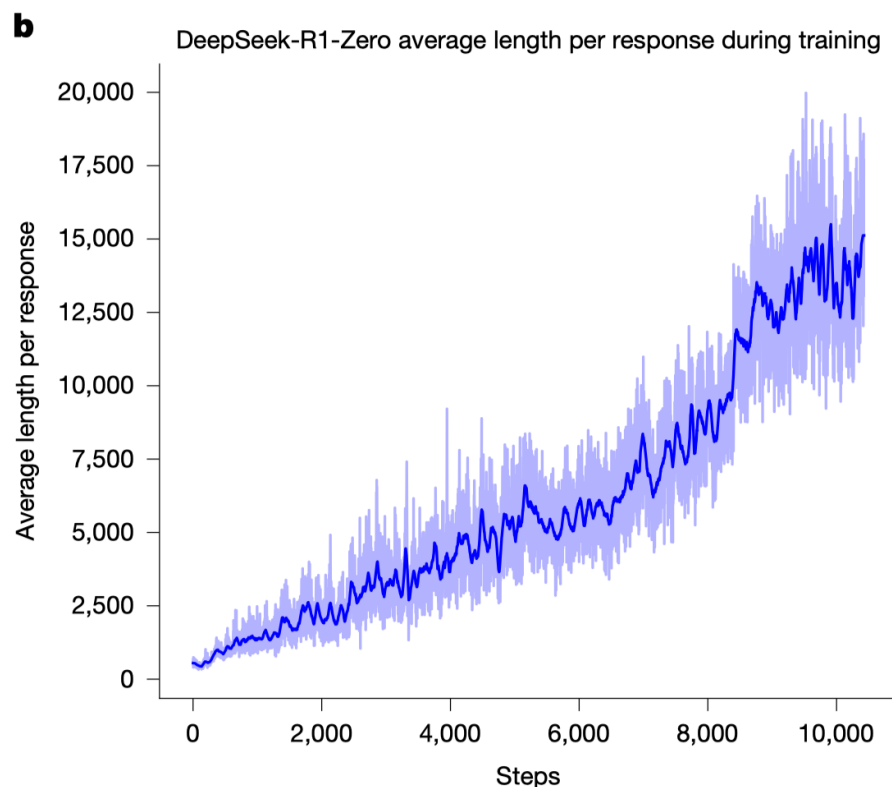
Reasoning = World Model + Search Strategy



Flexible and intuitive search

Flexibility

RL as meta-search: searching the search strategy



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a + x}})^2 = x^2 \Rightarrow a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \Rightarrow a^2 - 2ax^2 + (x^2)^2 = a + x \Rightarrow x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step by step to identify whether the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

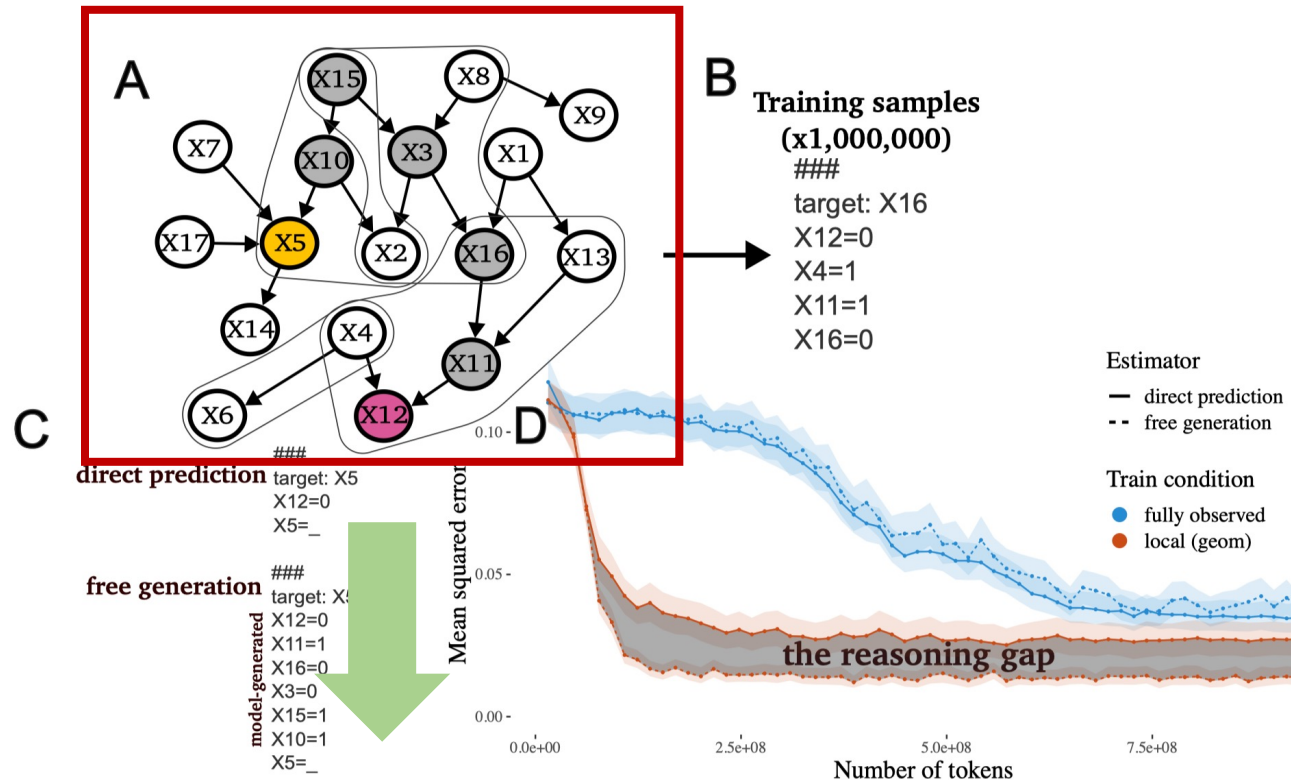
$$a - \sqrt{a + x} = x^2 \Rightarrow \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature 2025.

Reasoning = World Model + Search Strategy



Abstraction and Multimodality

Mental Models of Human

World Models

David Ha¹ Jürgen Schmidhuber^{2,3}



Jay Wright
Forrester
Father of system
dynamics
(1971)

Abstract Representations:

“The image of the world around us, which we carry in our head, is just a model. Nobody in his head imagines all the world, government or country. He has only selected concepts, and relationships between them, and uses those to represent the real system”

Which Representation to Predict?



**Abandon
generative models!**



"Modeling the world for action by **generating pixel is as wasteful** and doomed to failure..."

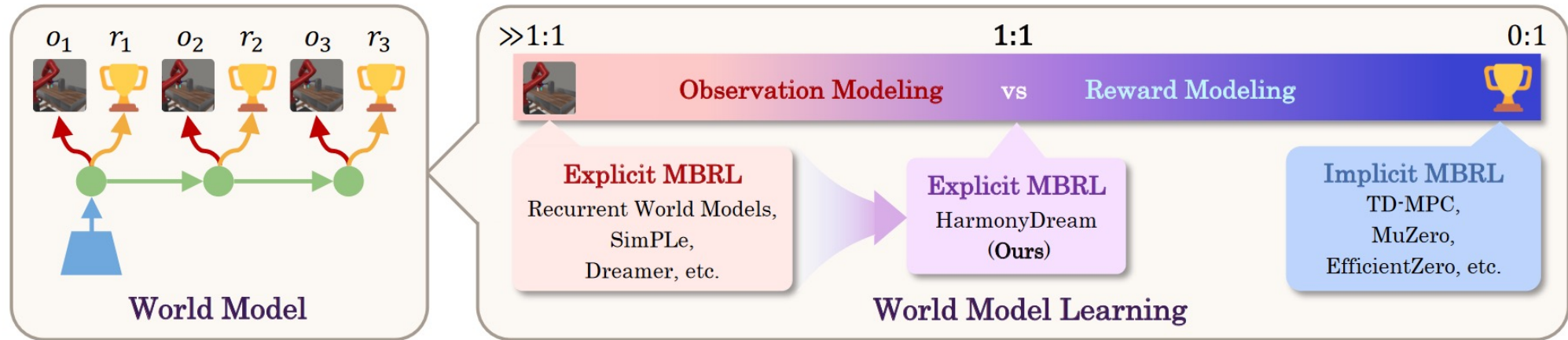
"It's much more desirable to generate **abstract representations** of those continuations that **eliminate details in the scene that are irrelevant** to any action we might want to take."

Raw Pixels vs. Abstraction

OpenAI. <https://openai.com/research/video-generation-models-as-world-simulators>

Yann LeCun. <https://twitter.com/ylecun/status/1758740106955952191>

Abstraction Matters: HarmonyDream (2024)



A Multi-task (Multi-modal) View of World Models

- **Observation Modeling:** how the environment transits and is observed

$$p(o_{t+1:T} \mid o_{1:t}, a_{1:T})$$

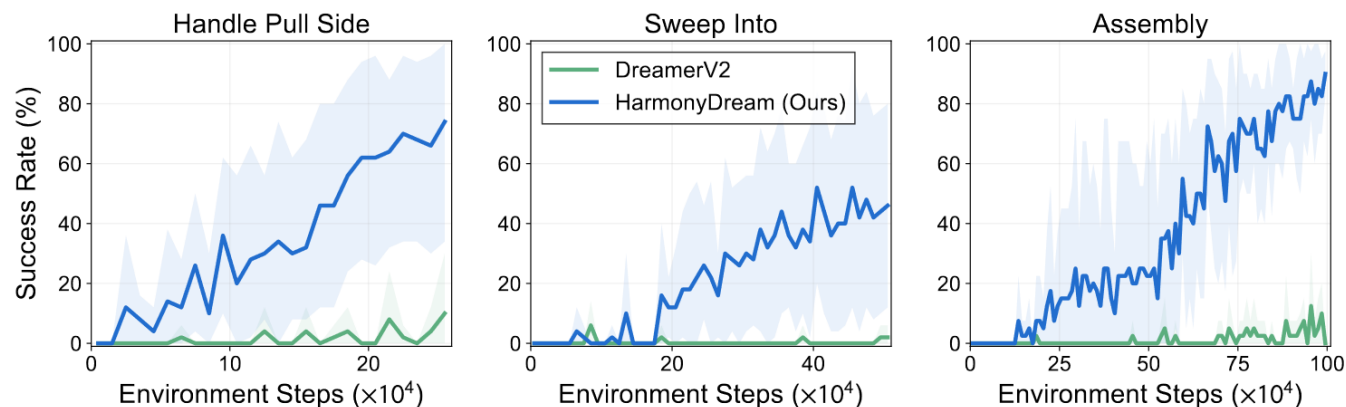
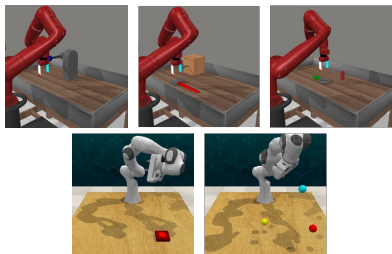
- **Reward Modeling:** how the task has been progressed

$$p(r_{t+1:T} \mid o_{1:t}, a_{1:T})$$

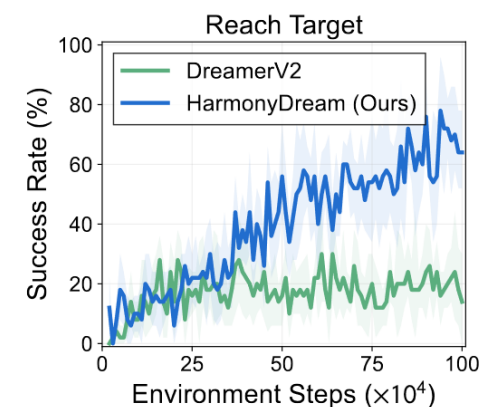
Adaptive weighting reward modeling leads to more task-oriented representations

Abstraction Matters: HarmonyDream (2024)

Robotic manipulation



(a) Meta-world

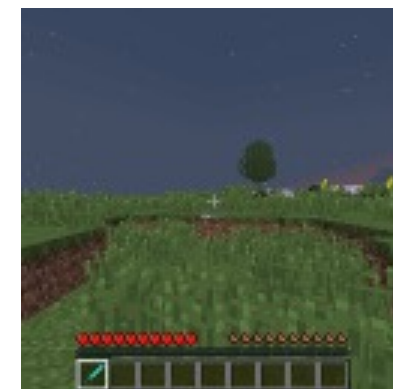
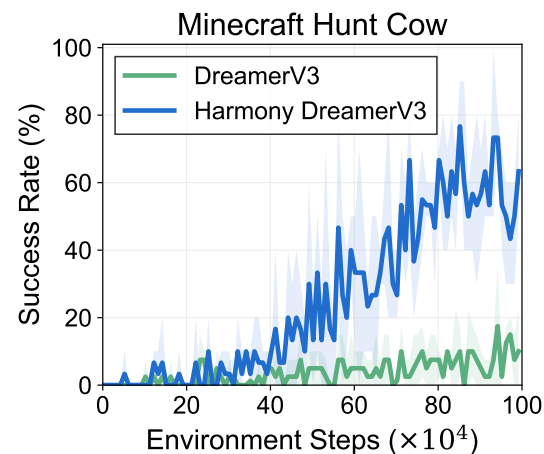
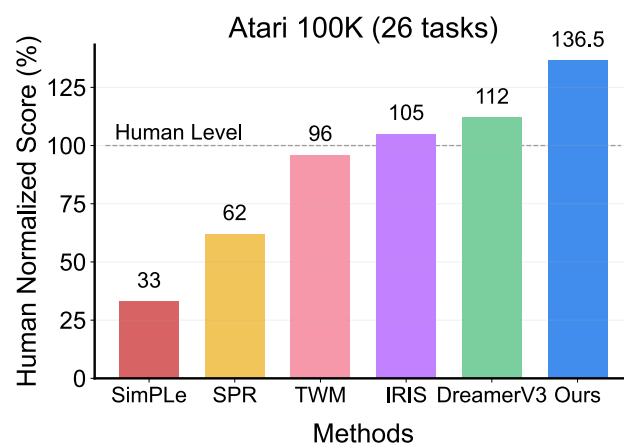
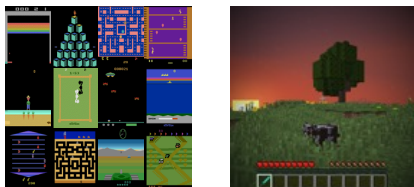


(b) RL Bench

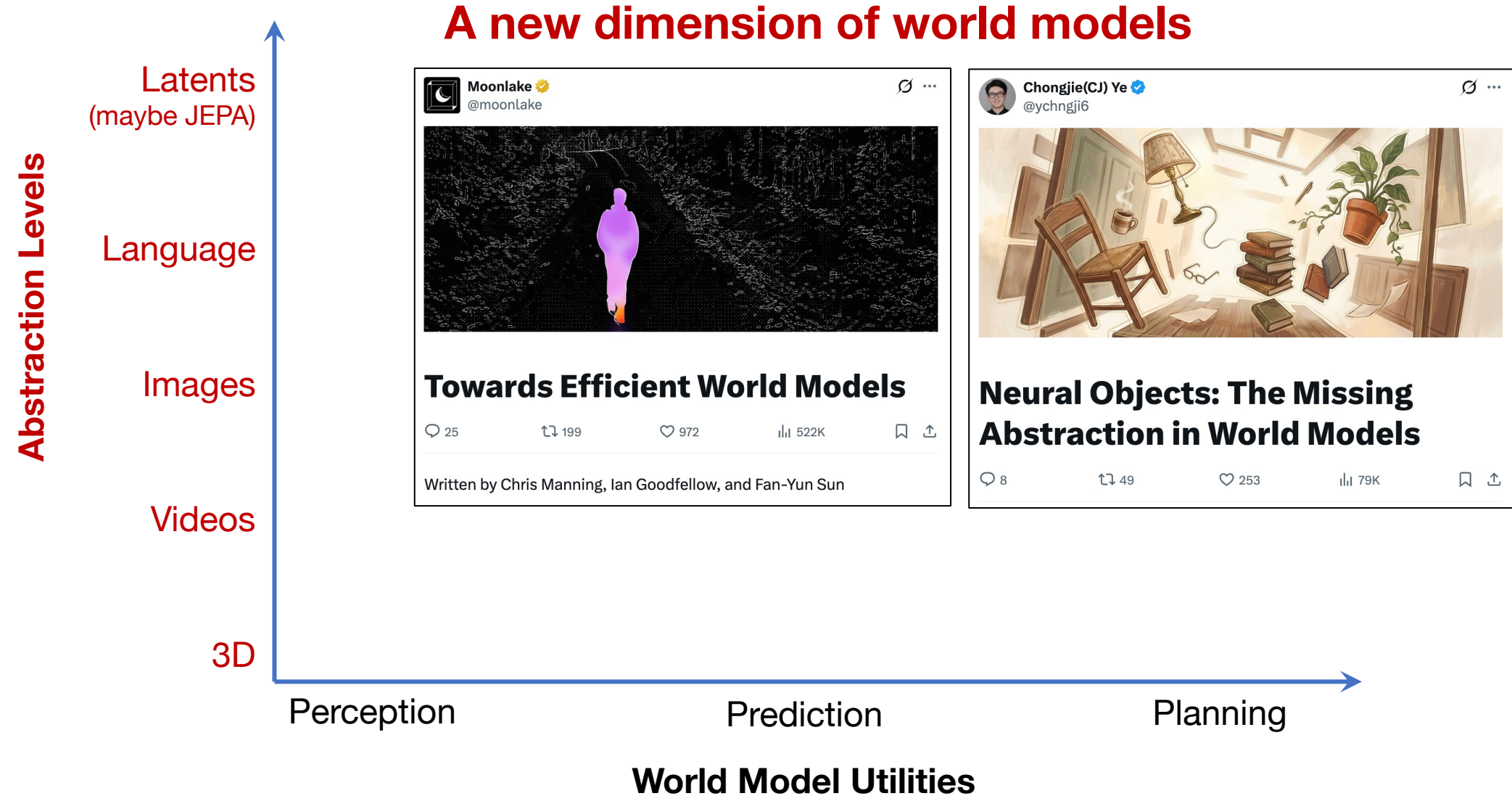
Distracted locomotion



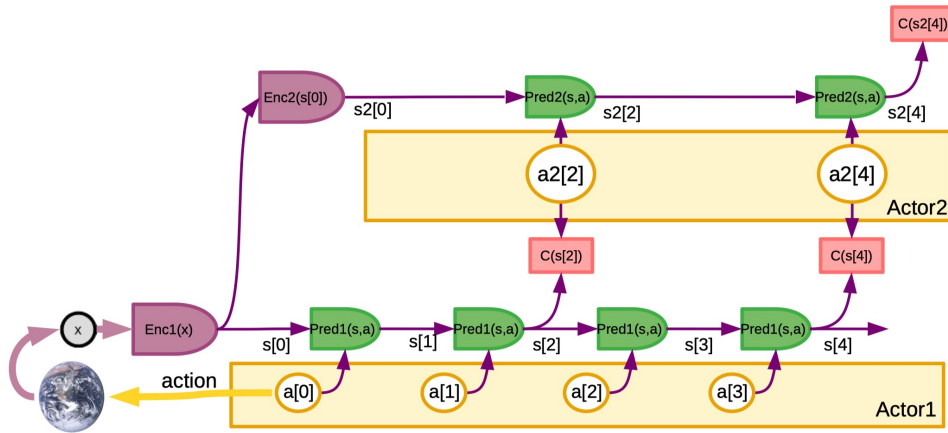
Video games



Different Levels of Abstraction



Different Levels of Abstraction



We have no objective signals in the wild!
How to learn **abstract latent embedding**
at scale?

JPEA?



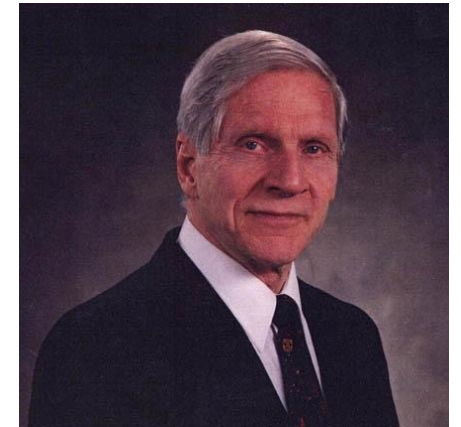
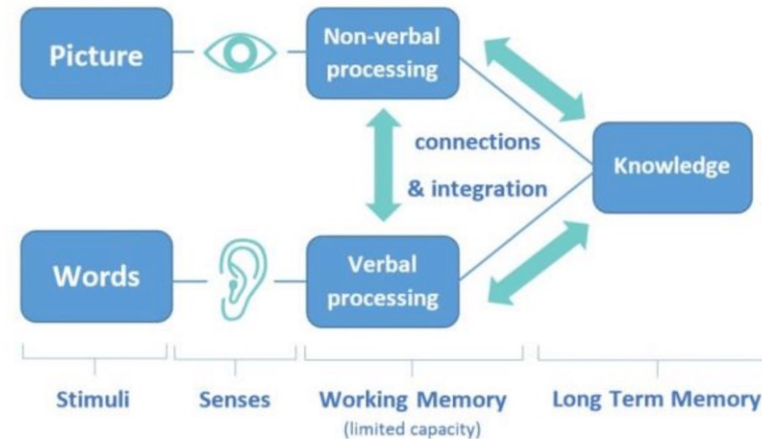
Approximate by

Multimodal world models
Visuals (concrete) + **Languages** (abstract)

Dual-Coding Theory of Human Cognition

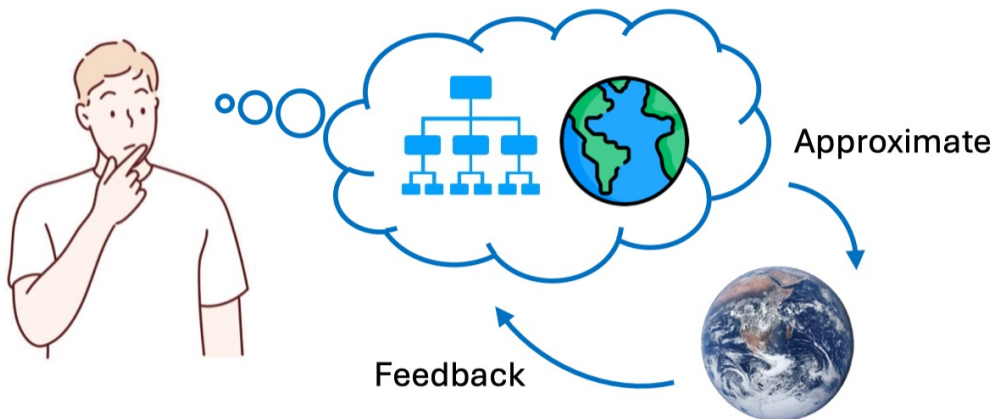
Mind processes information along two different channels: verbal and nonverbal.

Allan Paivio's Dual-Coding Theory



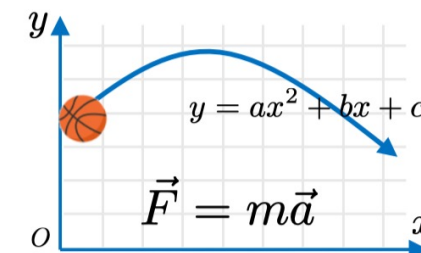
Allan Paivio
Professor of psychology

World Model: Mental Model of the World



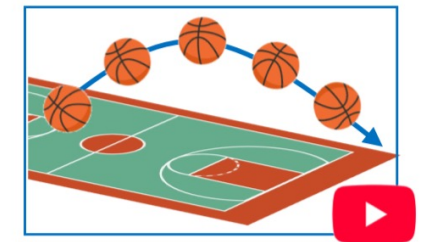
Dual Representations of World Knowledge

Verbal/Symbolic Knowledge



Dislike in Daily Life

Visual/Imagery Knowledge

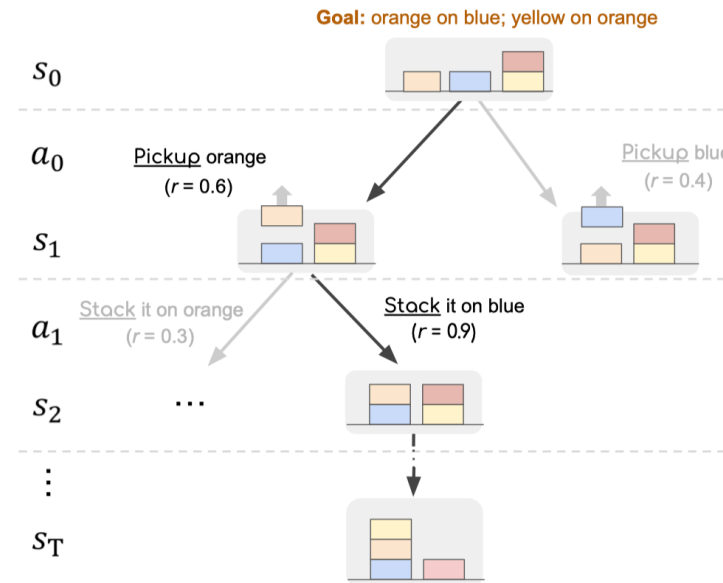


Prefer in Daily Life

Language as Abstract Representation of World States

Advantage of language:

- ✓ **Semantic Abstraction**
(discarding irrelevant details)
- ✓ **Temporal Abstraction**
(learning hierarchical policies)
- ✓ **Interpretability**
- ✓ **Prior knowledge from LLMs**



TextWorld

Welcome!

You are in the middle of the room. Looking around you, you see a diningtable, a stove, a microwave, and a cabinet.

Your task is to:
Put a pan on the diningtable.

> goto the **cabinet**

You arrive at the cabinet. The cabinet is closed.

> open the **cabinet**

The cabinet is empty.

> goto the **stove**

You arrive at the stove. Near the stove, you see a pan, a pot, a bread loaf, a lettuce, and a winebottle.

> take the **pan** from the **stove**

You take the pan from the stove.

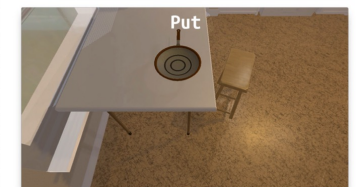
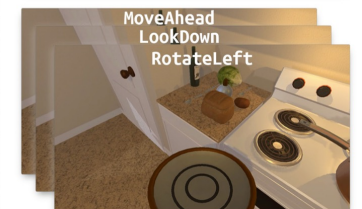
> goto the **diningtable**

You arrive at the diningtable.

> put the **pan** on the **diningtable**

You put the pan on the diningtable.

Embodied



Reasoning with Language Model is Planning with World Model. EMNLP 2023.

Mohit Shridhar, et al. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. ICLR 2021.

Chain-of-Thought Reasoning with **Verbal** World Models

World modeling as an inherent capability rather than a separate model.

During CoT reasoning, LLM/VLMs **explore, reflect, and backtrack** within the structured solution space, guided by their world knowledge.

Mathematical Reasoning

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...

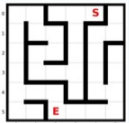
$$(\sqrt{a - \sqrt{a + x}})^2 = x^2 \Rightarrow a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \Rightarrow a^2 - 2ax^2 + (x^2)^2 = a + x \Rightarrow x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Puzzle Solving



<think>From S at [452, 59], the only available move is **West** to [359, 59]

Travel Planning

Task: Plan a trip with a budget of \$1,700.

<think>STATE: Initial Budget: \$1700, Spent: \$0.

Day 1: transportation: PENDING...Available:

Flight F3573659: \$474

ACTION: Plan day 1 transportation. Select Flight F3573659.

NEXT STATE: After this action, you've spent \$474, leaving \$1226 from your \$1700 budget...

State/Observation

Action

Everyday Activity Planning

Goal: Cooking tomato and eggs



<think>... **ACTION:** Cook the eggs in the pan
STATE: This action changes the state of the eggs from liquid to a partially cooked state. The eggs are now partially cooked and are being transformed into curds. The heat from the pan cooks the eggs, and the stirring action ensures even cooking...

Planning with reasoning using vision language world model.

Revisiting the necessity of lengthy chain-of-thought in vision-centric reasoning generalization.

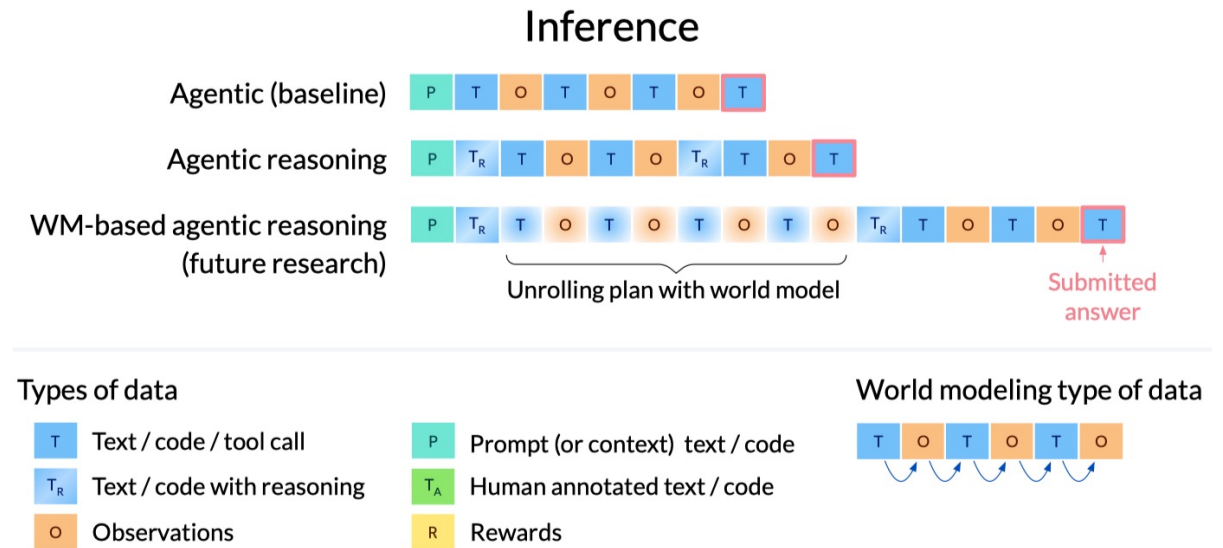
DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning.

Agent learning via early experience.

Chain-of-Thought Reasoning with **Verbal** World Models

World modeling as an inherent capability rather than a separate model.

During CoT reasoning, LLM/VLMs **explore, reflect, and backtrack** within the structured solution space, guided by their world knowledge.



CWM: An Open-Weights LLM for Research on Code Generation with World Models

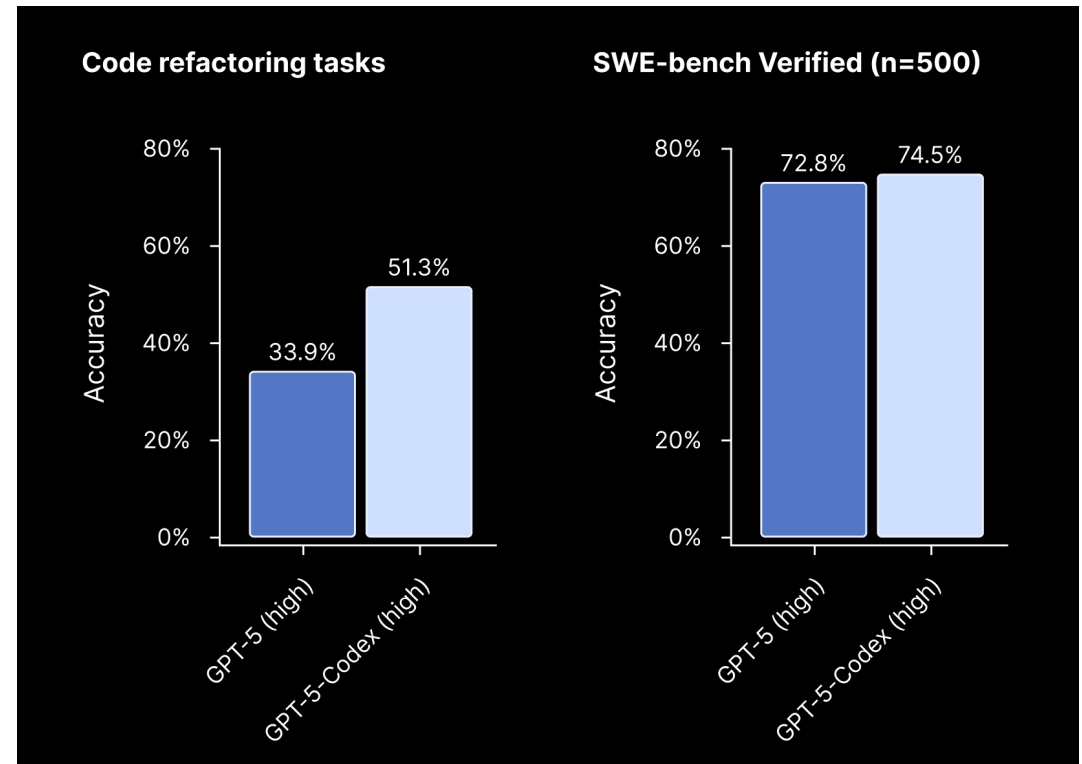
Expert Performance in Formal and Abstract Domains

Article | [Open access](#) | Published: 17 September 2025

DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning



Mathematics and Coding

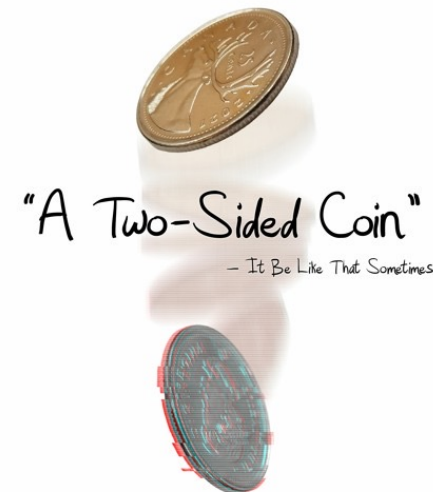


Grounding Also Matters: Visual Generation for Reasoning (2026)



Visual Generation Unlocks Human-Like Reasoning through Multimodal World Models

Jialong Wu^{1,2,*}, Xiaoying Zhang^{2,†}, Hongyi Yuan², Xiangcheng Zhang^{1,2,*}, Tianhao Huang¹, Changjing He¹, Chaoyi Deng^{1,2,*}, Renrui Zhang², Youbin Wu², Mingsheng Long^{1,†}

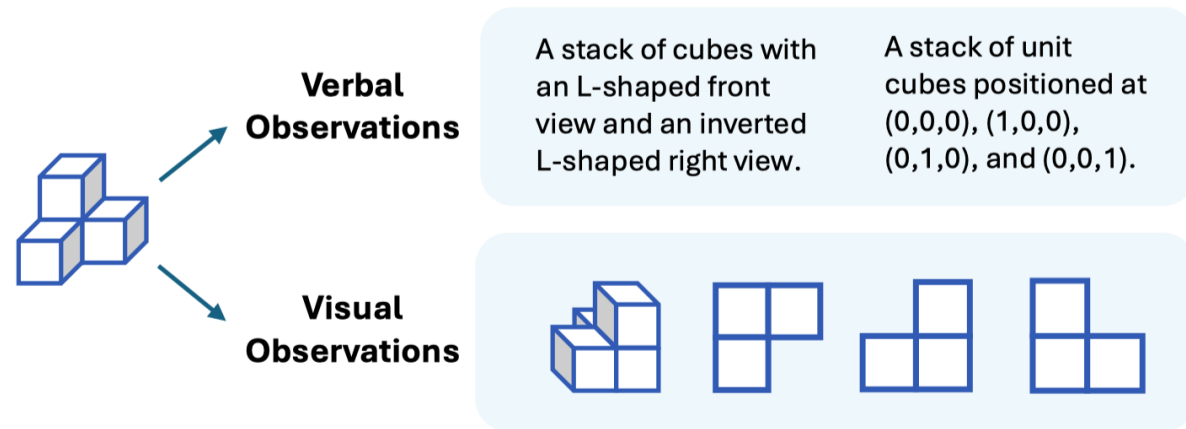


A perspective on:

1. What is **multimodal world models** and why do we need it?
2. When and how does **visual generation benefit reasoning**?

Multiple Observations of the World

Underlying world states can be observed through **multiple modalities with different focus**.



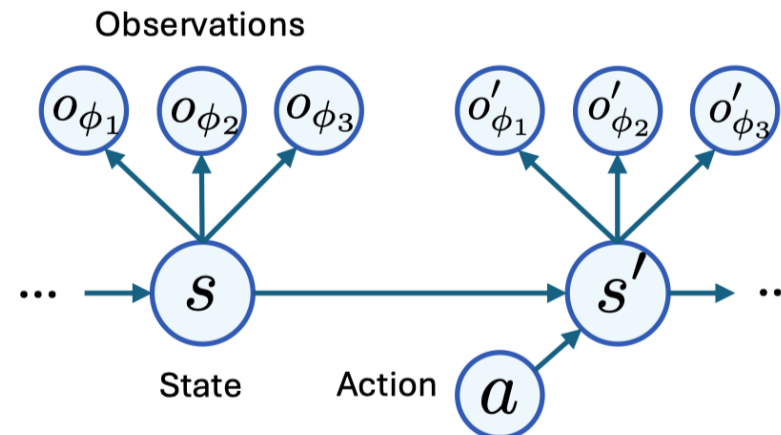
Multimodal world models approximate MOMDP:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \Phi, \mathcal{O}_\phi, e_\phi)$$

different observations (views)

parameterized by ϕ

Multi-Observable Markov Decision Process



Atomic Capabilities of World Models

World Reconstruction: reconstructing the environment structure from a few observations

Perception: $\hat{s} = \text{enc}(o_{\phi_1}, \dots, o_{\phi_n}) \approx s$

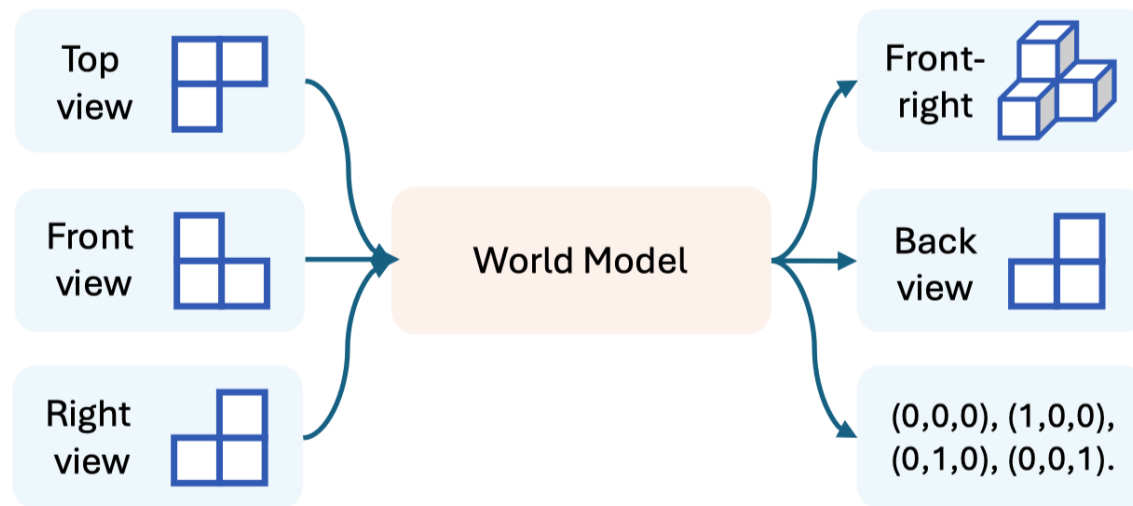
Decoding: $\hat{o}_{\phi_{n+1}} = \text{dec}(\hat{s}, \phi_{n+1}) \approx e_{\phi_{n+1}}(s)$

End-to-end



Novel view synthesis

$$p_{\theta}(o_{\phi_{n+1}} \mid o_{\phi_1}, \dots, o_{\phi_n})$$



Experience of **navigating** the world

Atomic Capabilities of World Models

World Simulation: simulating how the world evolves into the future

Perception: $\hat{s} = \text{enc}(o_{\phi_1}, \dots, o_{\phi_n}) \approx s$

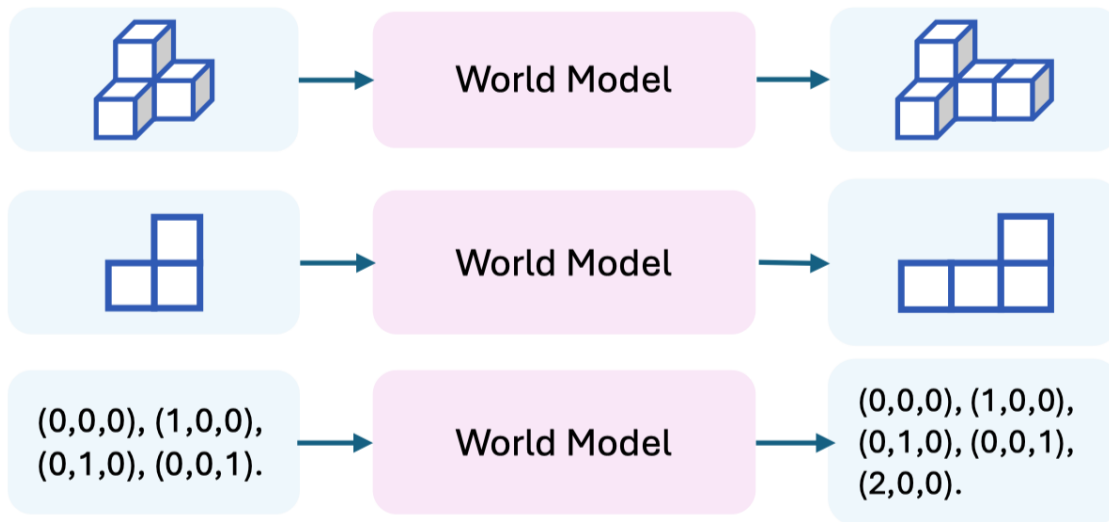
Prediction: $\hat{s}' \sim \text{pred}(\hat{s}, a)$

End-to-end



Future observation prediction

$$p_{\theta}(o_{t+1} \mid o_{\leq t}, a_{\leq t})$$



Experience of **interacting** with the world

Reasoning with World Modeling

Chain-of-thought reasoning

$$R = \tau_1, \tau_2, \dots, \tau_H$$

Intermediate steps

Refine



Explicitly maintain **evolving observations**,
generated by **world modeling** capabilities,
as **evidence of reasoning**

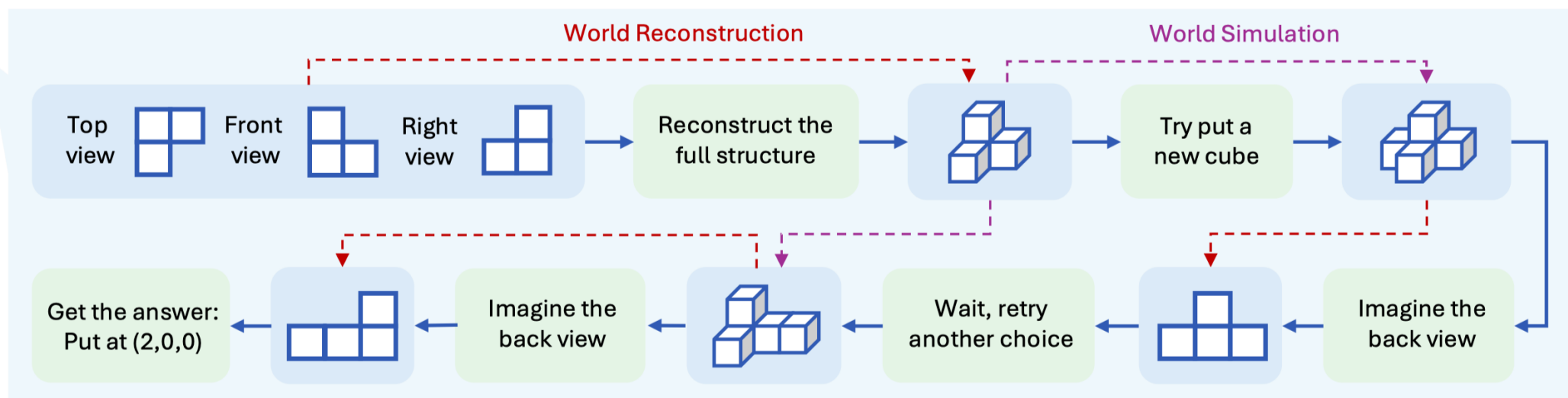
$$R = (\tau_1, o_1), (\tau_2, o_2), \dots, (\tau_H, o_H)$$

Logical reasoning step

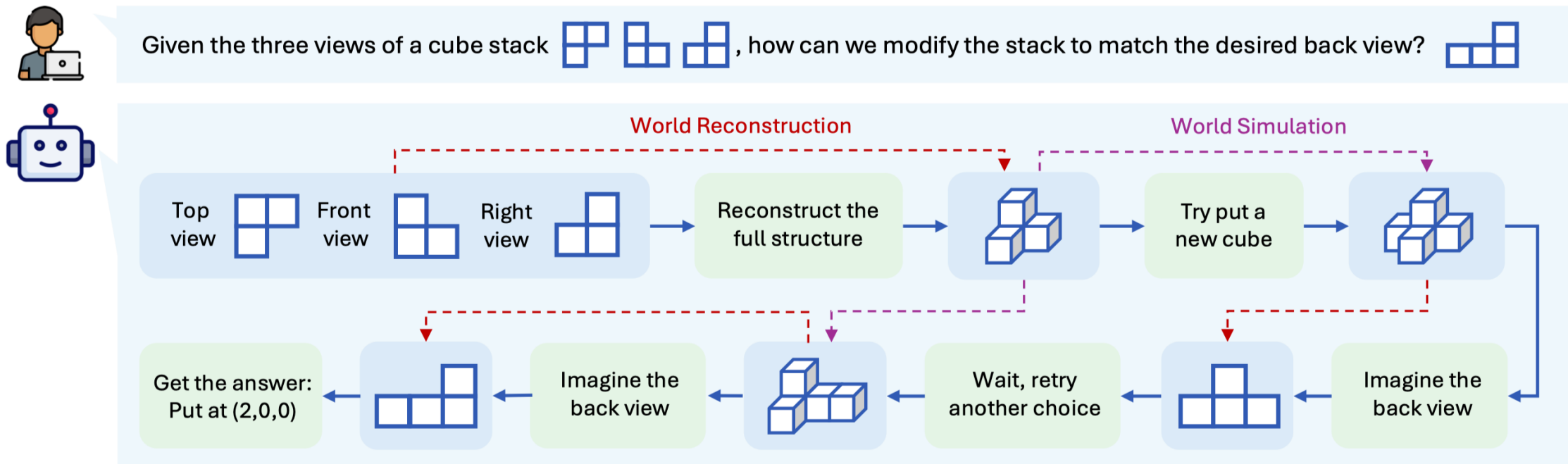
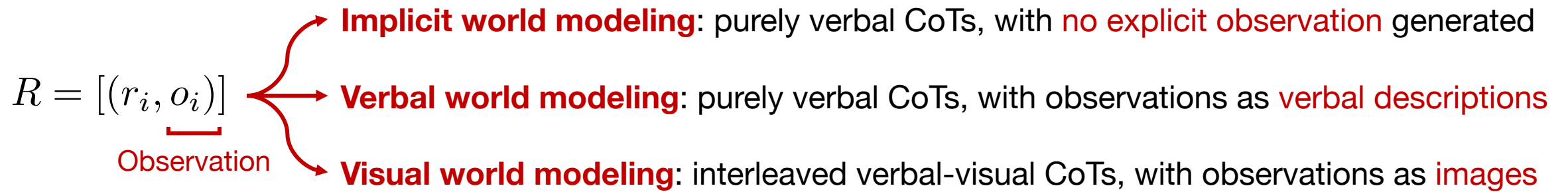
Observation



Given the three views of a cube stack   , how can we modify the stack to match the desired back view? 



Reasoning with World Modeling



Generated Interleaved CoTs

Real-World Spatial Reasoning

When you took the photo in Figure 1, where was the iron refrigerator located relative to you?



<think>First, let's analyze the images. [...] It's not visible in that initial view, so I need to change my perspective.



The refrigerator is not visible in this 90-degree left turn view, [...] To be thorough, I will also check the view to the right.



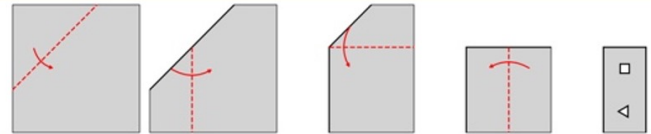
[...] My initial turn was 90 degrees left, but the refrigerator isn't at that exact angle. Let's try a smaller turn to the left.



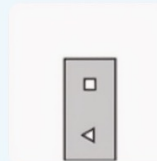
[...]

Paper Folding

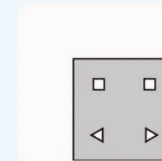
Analyze the image showing a folded paper with cutouts, calculate the number of triangle-left minus the number of triangle-right.



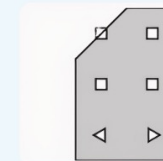
<think>First, zoom into the upper section's rightmost image [...]



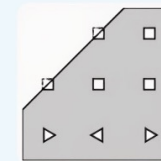
First, we will reverse the fourth and final fold. This was a vertical fold where [...]



Next, we reverse the third fold. [...]



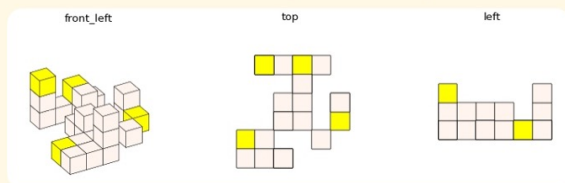
Continuing the process, we now undo the second fold. This was a vertical fold where [...]



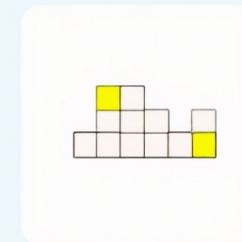
[...]

Cube 3-View Projection

How many cubes in yellow can possibly be seen from the front view? A. 2. B. All three other options are possible. C. 4. D. 1.



<think>To determine the number of yellow cubes visible from the front, we first need to construct the front view of the stack. By synthesizing the provided isometric, top, and left views, we can generate the required front view.



Examining the generated front view, we can directly count the number of visible yellow cubes. There are 2 yellow cubes visible. Therefore, the correct answer is 2, which corresponds to option A.</think> A

The Visual Superiority Hypothesis

Informativeness

(Information theoretical analysis)



Prior Knowledge

(Transfer learning analysis)



Verbal representations suffer from **ambiguity** and **representational bottlenecks**



Humans naturally **represent and communicate** certain knowledge visually



The Visual Superiority Hypothesis

In multimodal reasoning tasks grounded in the **physical world**, visual generation as a world model yields representations that are more **informative** and **knowledge-rich** than those produced by verbal world models.

Key Findings

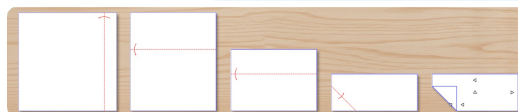
1. Visual World Simulation Boosts Multimodal Reasoning
2. Visual World Reconstruction Boosts Multimodal Reasoning
3. Visual World Modeling is Unhelpful for Certain Tasks
4. Do UMMs Compromise Verbal Reasoning Capabilities and Bias Comparisons? No
5. RL Enhances Various CoTs, Yet Does Not Close the Gap.

VisWorld-Eval: Task Suite for Reasoning with Visual World Modeling

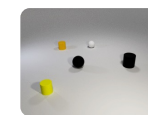
World Simulation

Paper Folding

Q: How many cutouts are there in the unfolded paper? A: 15

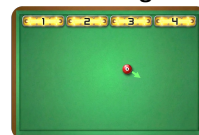


Multi-Hop Manipulation



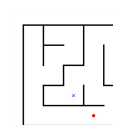
Q: Starting with the initial arrangement, perform the following:
1. Place a red cylinder to the left of the black cylinder.
2. Swap the colors of the orange cylinder and the black cylinder.
After these operations, what is to the left of the orange cylinder?
A. black sphere, B. white sphere, C. yellow cylinder, D. red cylinder.

Ball Tracking



Q: Given a red point-mass ball that moves at constant speed, reflects perfectly off solid walls, and follows the initial direction indicated by a green arrow, determine which numbered hole at the top it will enter first. A: 1

Maze



Q: Navigate the maze from the red dot to the blue X.
A: (4, 5), (5, 5), (5, 4) ...

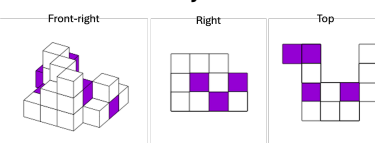
Sokoban



Q: Guide the player to push the box onto the goal position.
A: Down, Right, Down, ...

World Reconstruction

Cube 3-View Projection

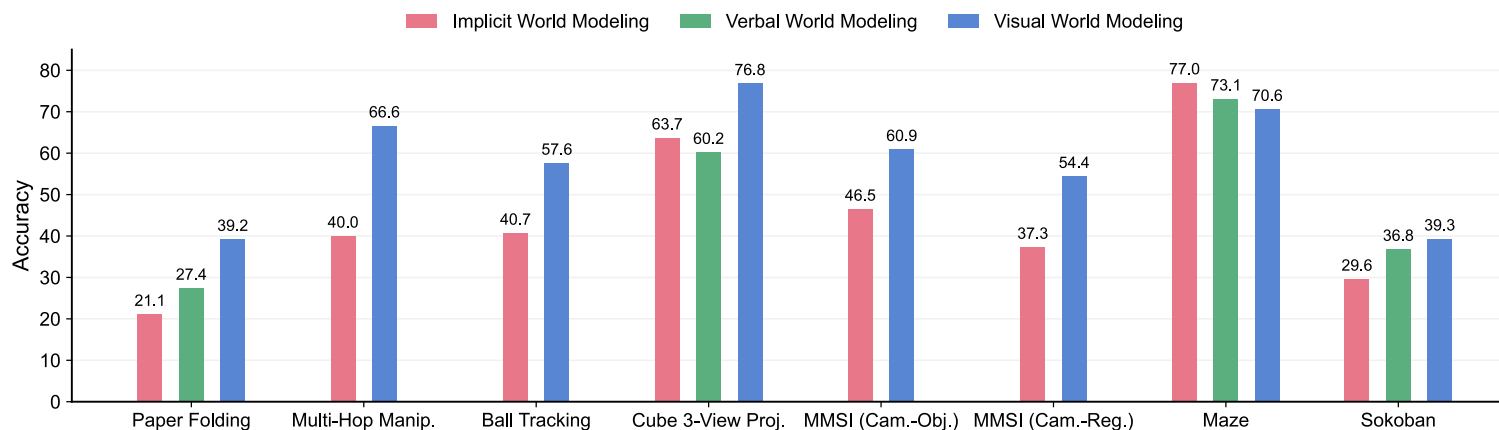


Q: How many cubes in dark violet can possibly be seen from the back view?
A. 0, B. 2, C. 3, D. 9.

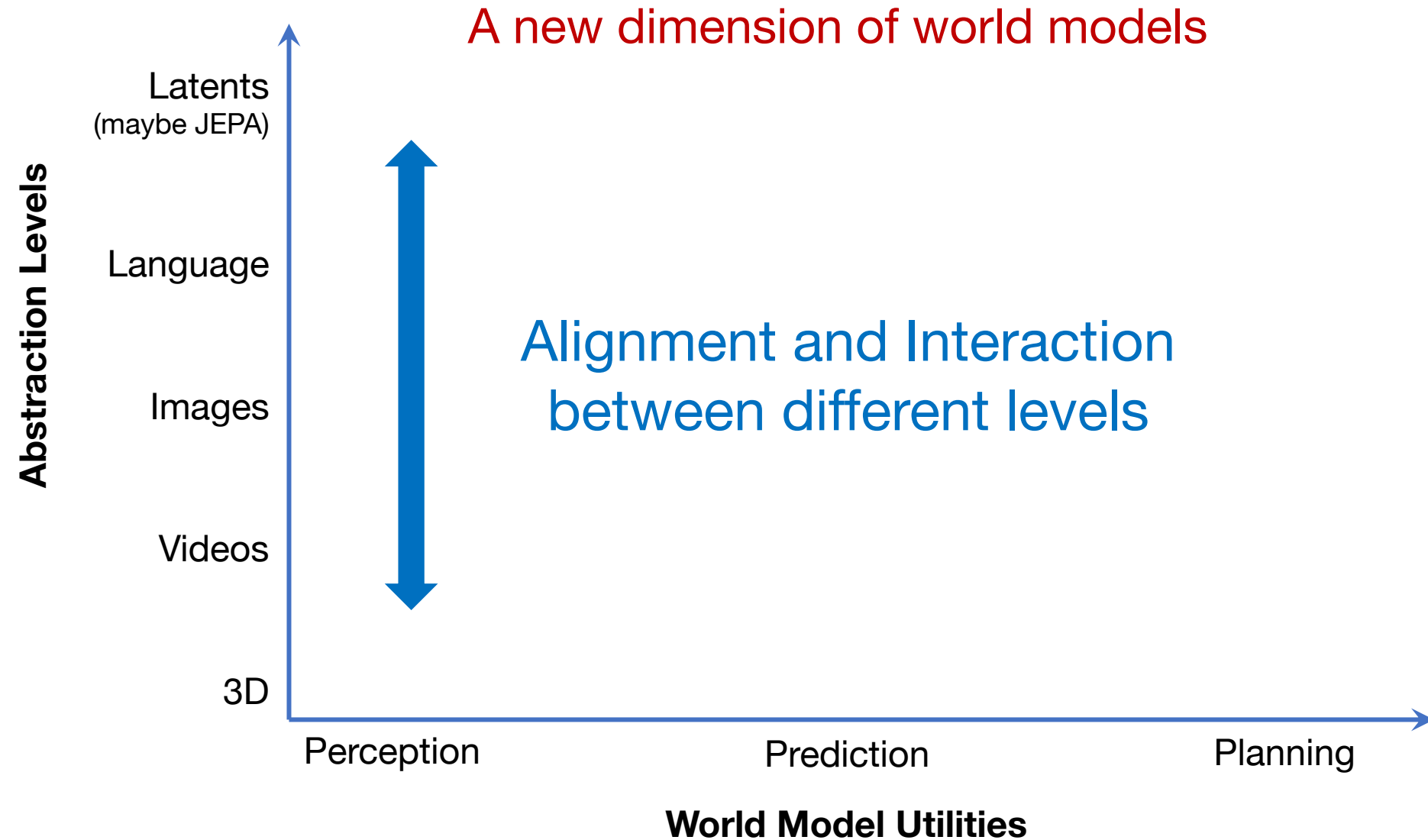
Real-World Spatial Reasoning



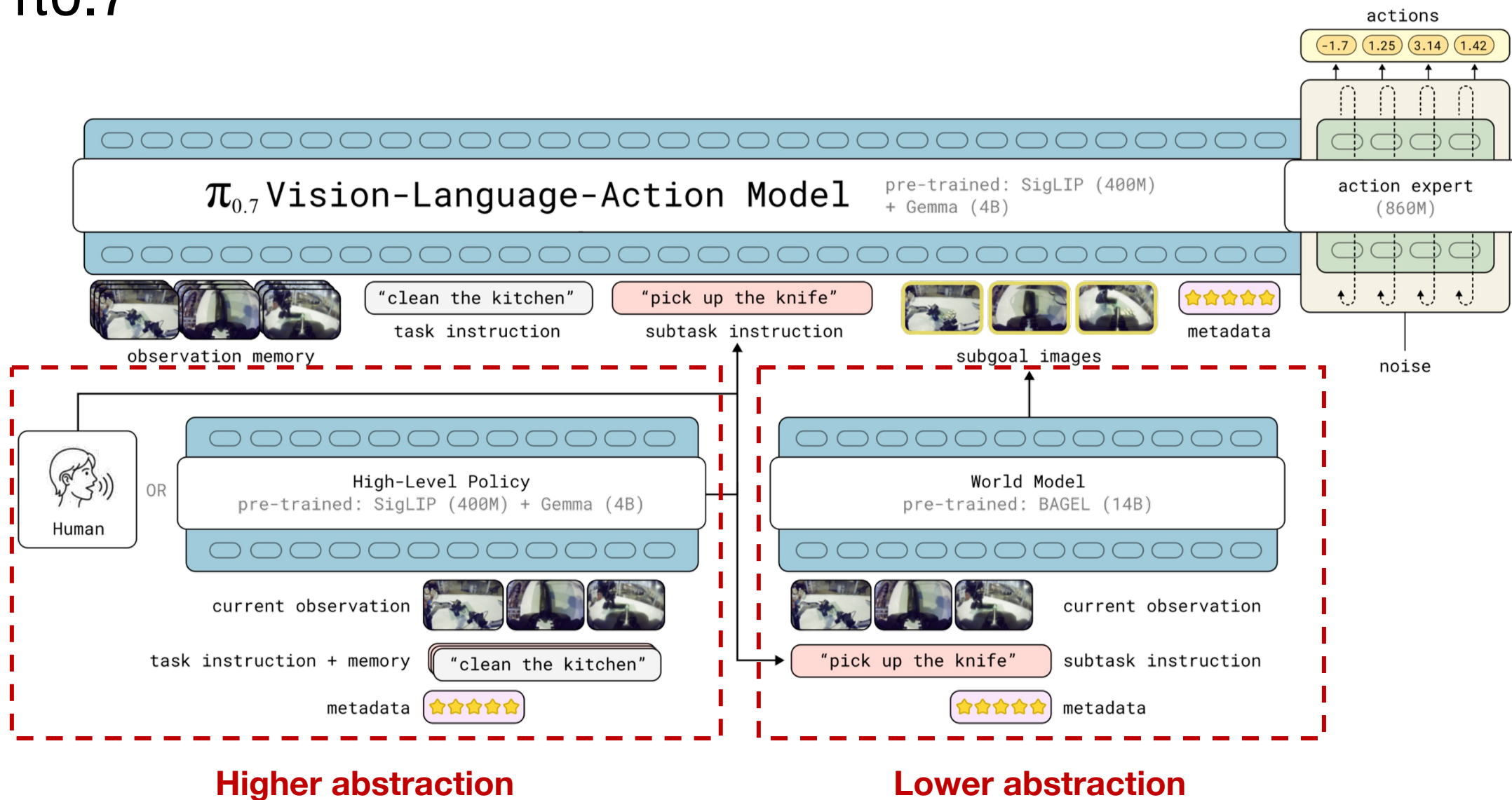
Q: Which direction is the black door relative to me when I am taking Image 2?
A. Behind, B. Left, C. Front, D. Right



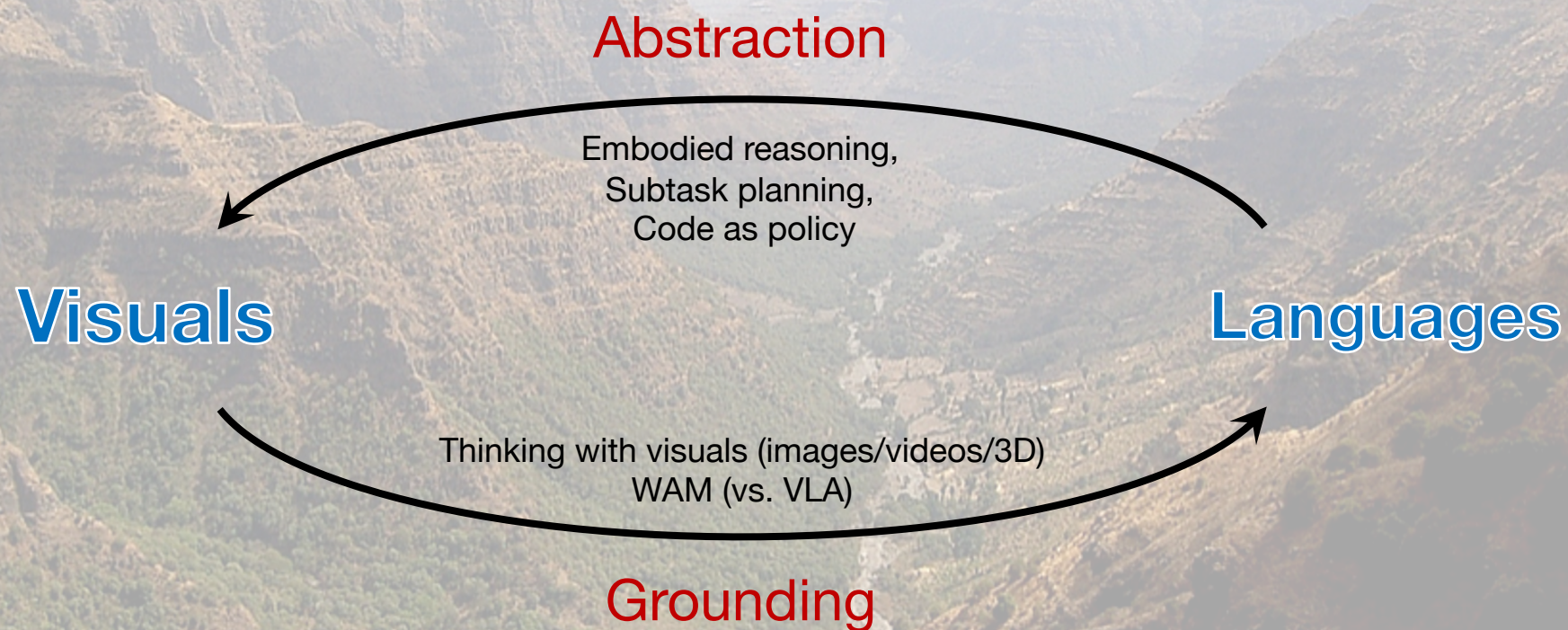
Different Levels of Abstraction



$\pi_{0.7}$

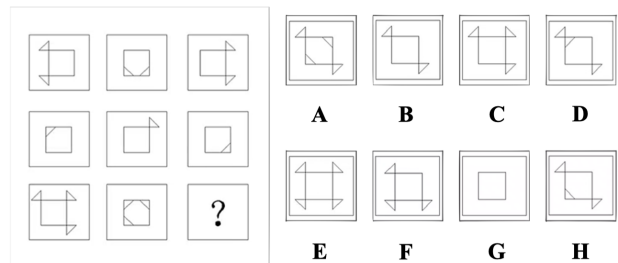


The Gap Remains Large



Baby Performance in Physical and Spatial Reasoning

Struggle with tasks grounded in basic physical and spatial intuition that even young children naturally master



Question: Based on the sequence and position of the other shapes, identify the pattern and determine the correct option for the question mark grid.



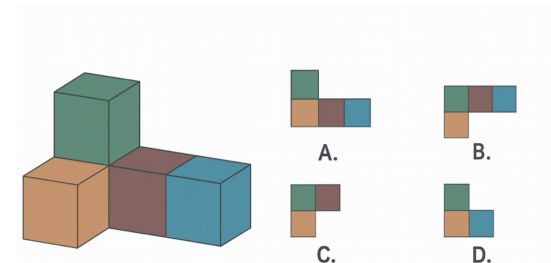
GPT-5

Answer: F. ✓



Human

Answer: F. ✓



Question: Which option is the correct top-down view of the object?



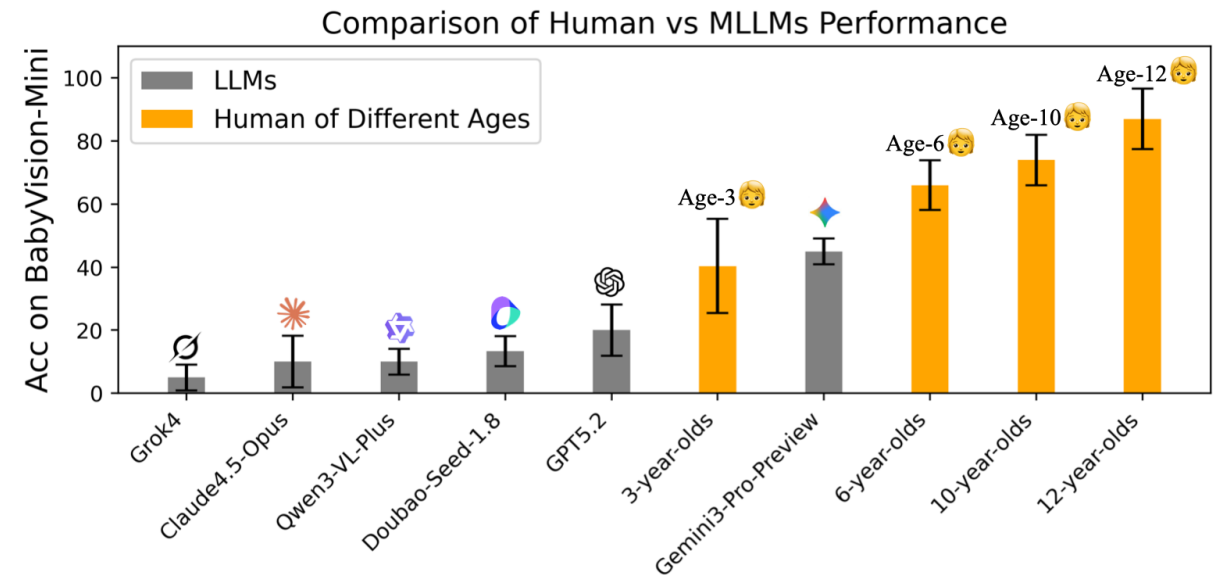
GPT-5

Answer: A. ✗



Human

Answer: B. ✓

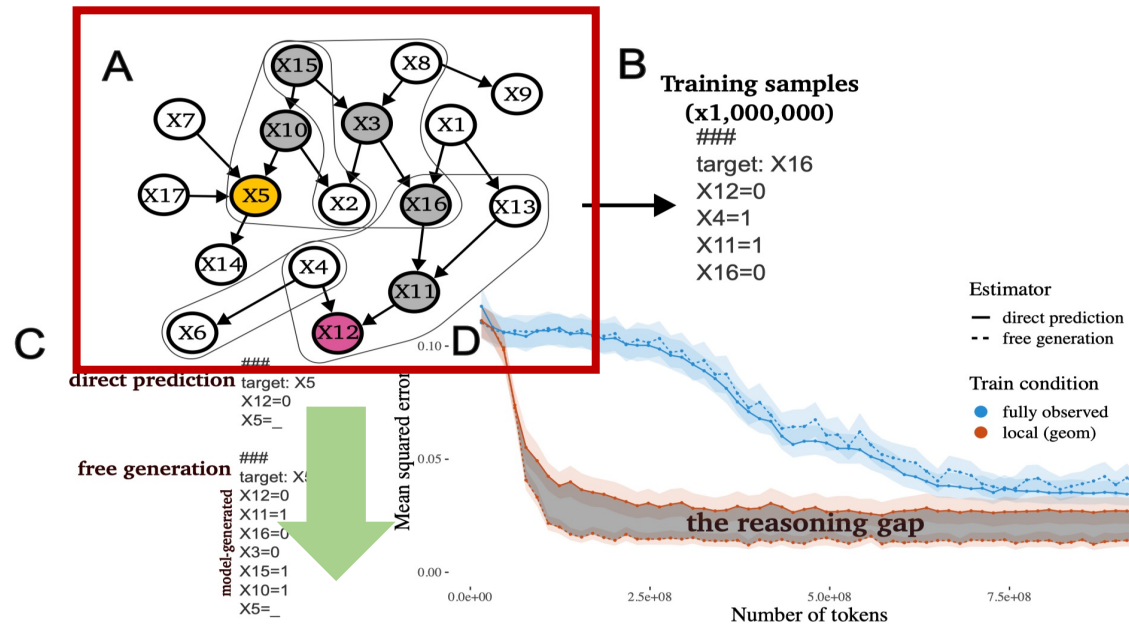


Has GPT-5 Achieved Spatial Intelligence? An Empirical Study

BabyVision: Visual Reasoning Beyond Language

Bridging World Models and Multimodal Reasoning

Reasoning = World Model + Search Strategy



Abstraction and multimodality



Flexible and intuitive search

Unified Model

Thank You for Your Attention!