Foundation Models for Decision Making

Problems, Methods, and Applications

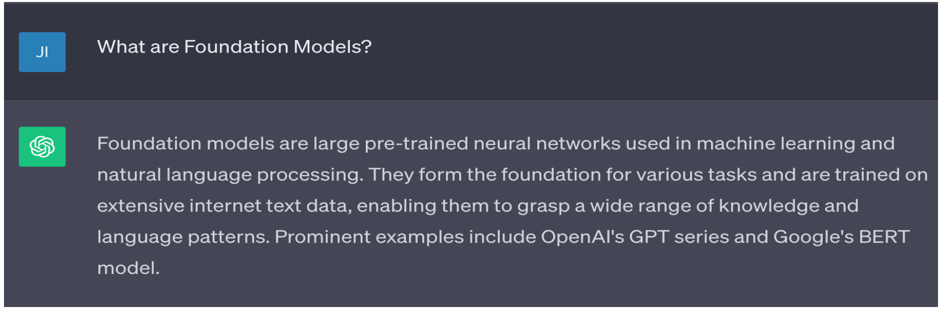
Sherry Yang

Berkeley UNIVERSITY OF CALIFORNIA

Google DeepMind

Machine Learning Advances in Vision and Language

Text to image / video



Language generation

Behind These Advances: Foundation Models

JI

What are Foundation Models?

Foundation models are large pre-trained neural networks used in machine learning and natural language processing. They form the foundation for various tasks and are trained on extensive internet text data, enabling them to grasp a wide range of knowledge and language patterns. Prominent examples include OpenAI's GPT series and Google's BERT model.

Response from GPT-4

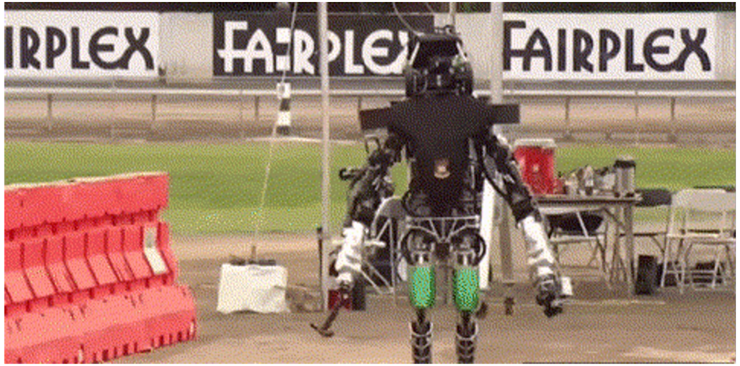
Modeling the Data is Not Enough

Issue: Not enough data

- ➢ Scientific discoveries
- ➢ Rare events, safety

Issue: Want better than data

- ➢ Failed robot executions
- ➢ Faster programs



[1] **Yang** et al. Foundation Models for Decision Making. arXiv 2023.

Promises of Sequential Decision Making

Issue: Not enough data

Solution: Collect more data

Issue: Want better than data

Solution: Optimize actions

[1] Sutton and Barto. Reinforcement Learning: An Introduction.1999.

Promises of Sequential Decision Making

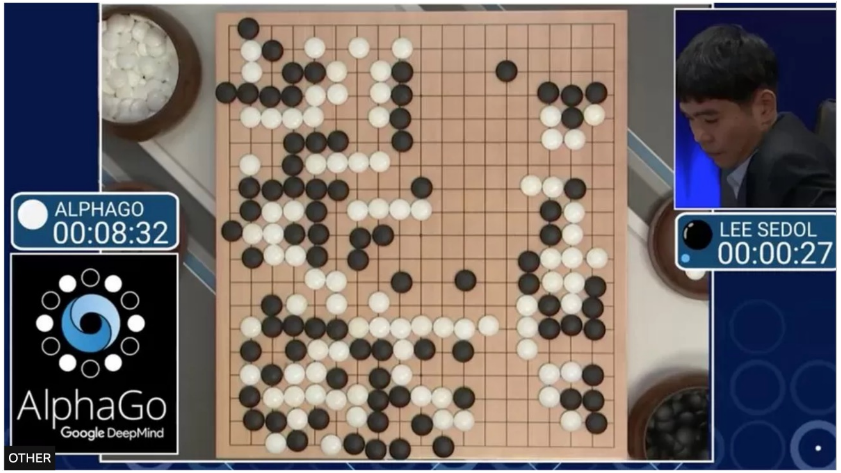
Issue: Not enough data

Solution: Collect more data

Issue: Want better than data

Solution: Optimize actions

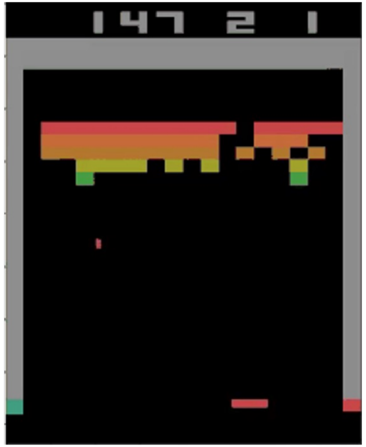
- ➢ Reinforcement learning
- ➢ Planning, search
- ➢ Control, optimization



Challenges of Sequential Decision Making

Solution: Collect more data

Challenge: Sample Efficiency



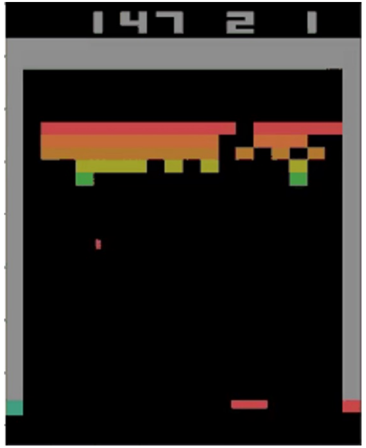
- ➢ RL: 38 days
- ➢ Human: mins

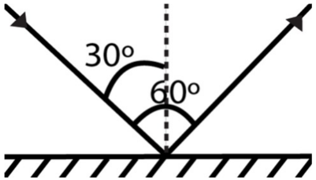
Solution: Optimize actions

Challenge: Generalization

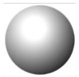
[1] Minh et al. Human-Level Control through Deep Reinforcement Learning. Nature 2015.
[2] Zhang et al. A Study on Overfitting in Deep Reinforcement Learning. arXiv 2018.

Sequential Decision Making Lacks Broad Knowledge





“Bounce the ball back.”

 and 

Physics **Language** **Vision**

How Foundation Models Acquire Broad Knowledge

Representation Learning

- ➢ Contrastive learning (SimCLR, CLIP)
- ➢ Denoising autoencoding (BERT, MAE)

Reasoning

Input

Intermediate steps

Output

Internet Data

WIKIPEDIA

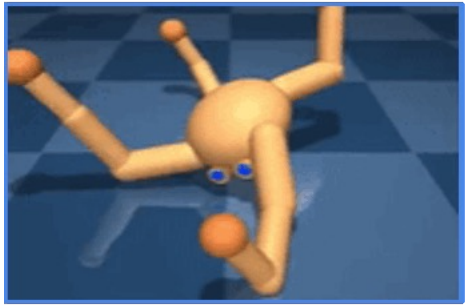
Common Crawl

[1] Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. PMLR 2020.
[2] Radford et al. Learning Transferable Visual Models From Natural Language Supervision. PMLR 2021.
[3] Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.
[4] He et al. Masked Autoencoders are Scalable Vision Learners. CVPR 2022.
[5] Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.
[6] Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 2022.

Today's Talk: Foundation Models for Decision Making

Representation Learning

From suboptimal data

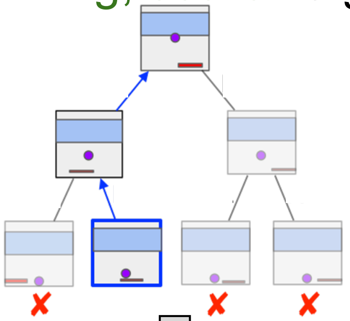


[ICML21, NeurIPS21, ICLR22, ICML22]

Reasoning

State

Planning, search algos



Action

[NeurIPS22]

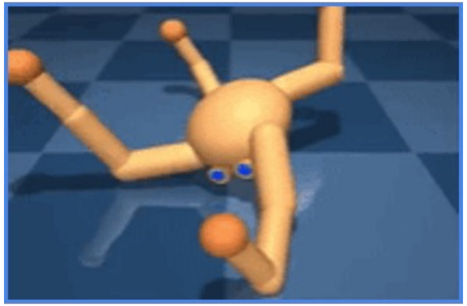
Internet Data

[NeurIPS23, arXiv23, arXiv23]

Today's Talk: Foundation Models for Decision Making

Representation Learning

From suboptimal data



[ICML21, NeurIPS21, ICLR22, ICML22]

Reasoning

Input

Planning, search algos

Output

[NeurIPS22]

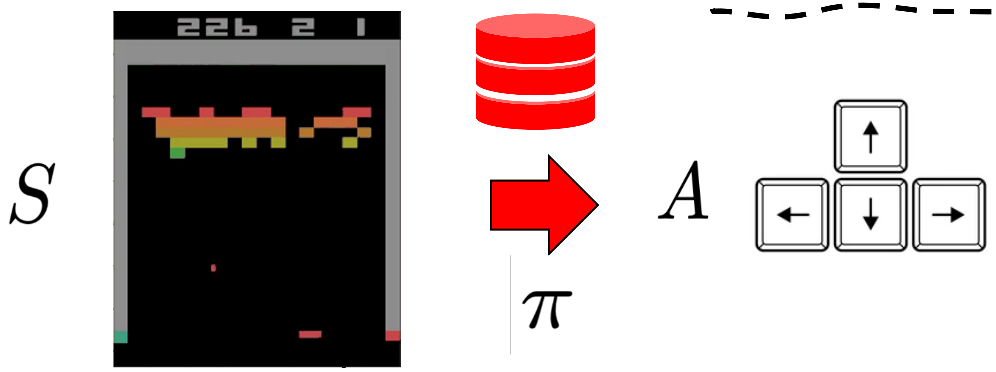
Internet Data

[NeurIPS23, arXiv23, arXiv23]

Learning from Expert Demonstrations

π_* Optimal policy

Imitation learning:

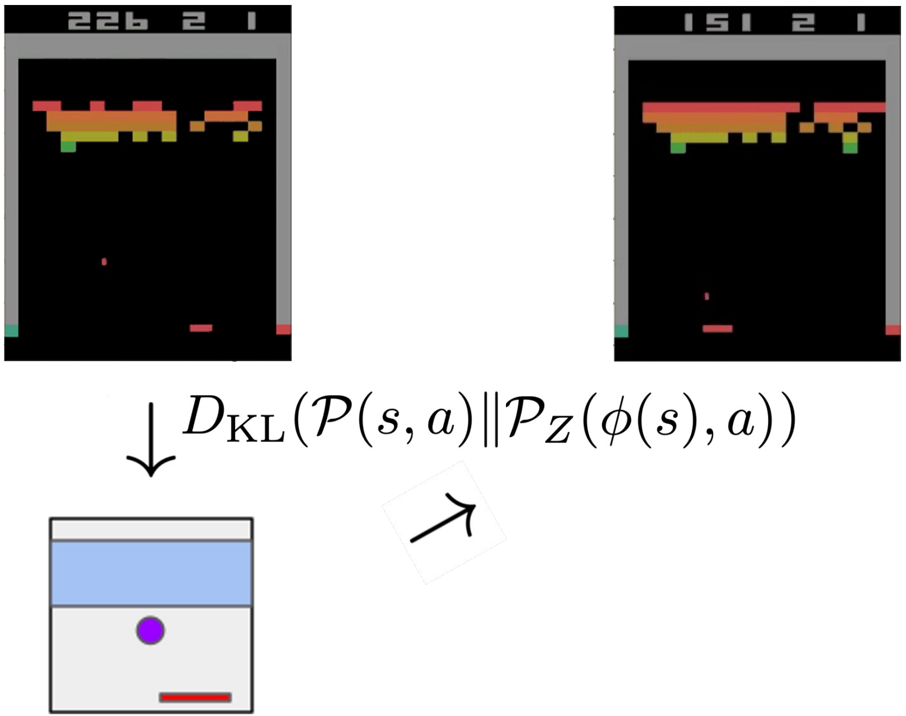


Representation Learning from Suboptimal Data

Suboptimal data

Pretraining

$\phi : S$



s s'

$\downarrow D_{\mathrm{KL}}(\mathcal{P}(s,a) \| \mathcal{P}_Z(\phi(s),a))$

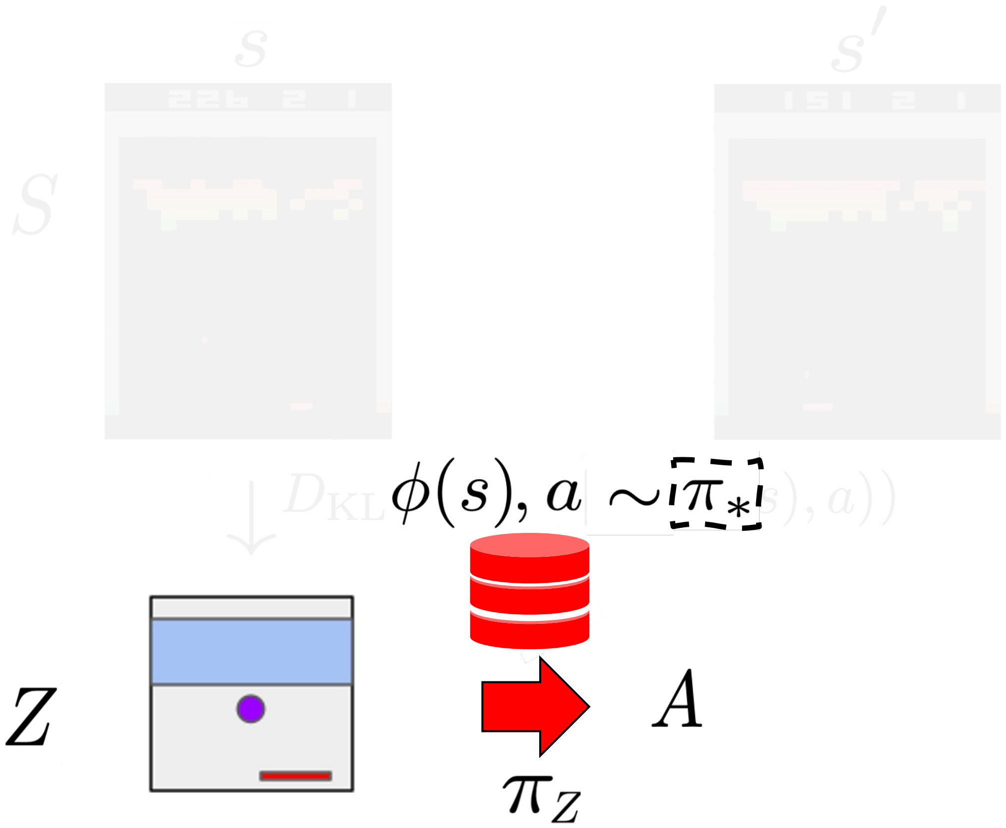
$\nearrow$

Z

[1] Nachum and **Yang**. Provable Representation Learning for Imitation. NeurIPS 2021.

Representation Learning from Suboptimal Data

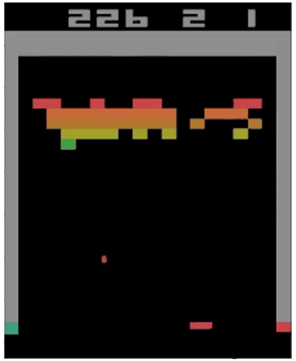
Imitation with representations:



[1] Nachum and **Yang**. Provable Representation Learning for Imitation. NeurIPS 2021.

Intuition: Why Representation Learning Helps

Imitation learning: S

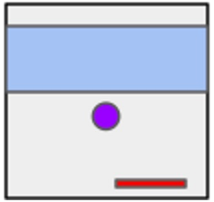


$\xrightarrow{\pi} A$

- Smaller hypothesis space.
- Need fewer expert demos.

$$\boxed{|Z| < |S|}$$

Imitation with representations: Z

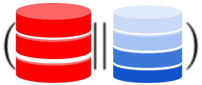


$\xrightarrow{\pi_Z} A$

[1] Nachum and **Yang**. Provable Representation Learning for Imitation. NeurIPS 2021.

Performance Difference with Representations

Theorem: For any expert policy **π***, representation ϕ, and policy **π$_Z$**, PerfDiff(**π$_Z$** , **π***) is bounded

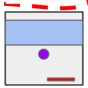
$\boxed{\mathrm{PerfDiff}(\pi_Z, \pi_*)} \leq (1 + D_{\chi^2}($  $)^{\frac{1}{2}}) \cdot \boxed{\epsilon_{\mathrm{R,T}}} + C\sqrt{\frac{1}{2}\underbrace{\mathbb{E}_{z\sim d_Z^{\pi_*}}[D_{\mathrm{KL}}(\pi_{*,Z}(z)\|\pi_Z(z))]}_{= \mathrm{const}(\pi_*, \phi) + \boxed{J_{\mathrm{BC},\phi}(\pi_Z)}}}$

Learning Goal

$\propto D_{\mathrm{KL}}(\mathcal{P}(s,a)\|\mathcal{P}_Z(\phi(s),a))$

Approx. dynamics

Sample complexity $\propto |Z|$

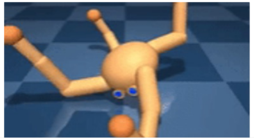
Downstream imitation in 

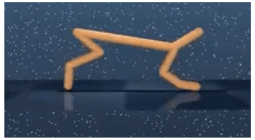
- Expect improvement when $\epsilon_{\mathrm{R,T}}$ and $|Z|$ are small.
- Vanilla BC corresponds to $\epsilon_{\mathrm{R,T}} = 0$ and $|Z| = |S|$.

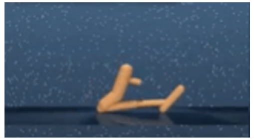
[1] Nachum and **Yang**. Provable Representation Learning for Imitation. NeurIPS 2021.

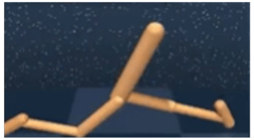
Empirical Results on Continuous Control

Suboptimal data





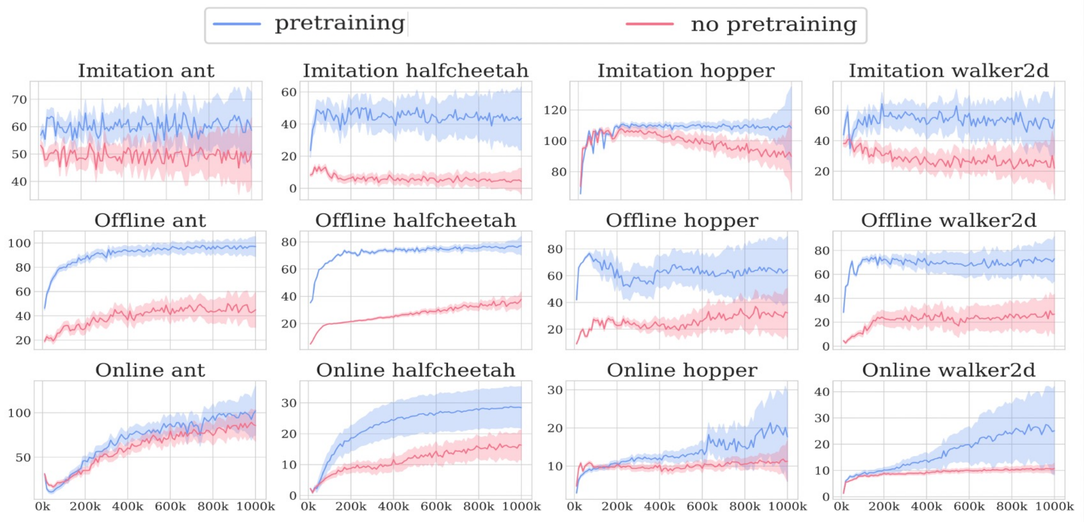




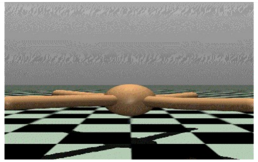
Imitation

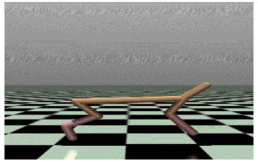
Offline RL

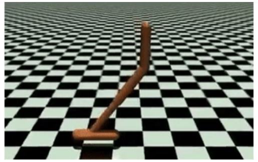
Online RL

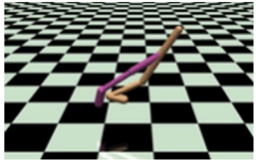


With representation





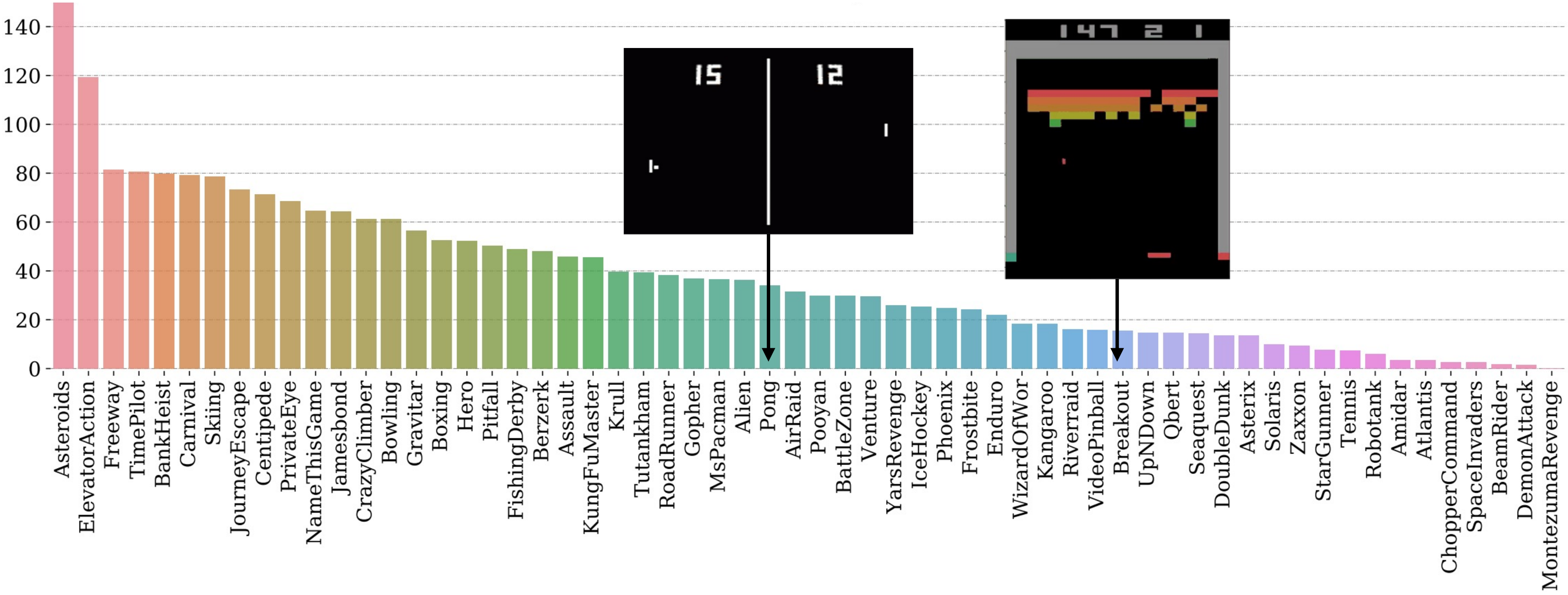




[1] **Yang** and Nachum. Offline Pretraining for Sequential Decision Making. ICML 2021.

Empirical Results on Atari Games

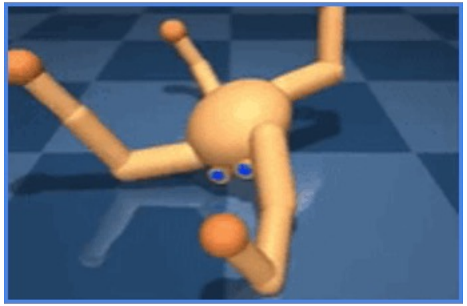
Improvement % over Behavioral Cloning (BC) without representation learning



[1] Nachum and **Yang**. Provable Representation Learning for Imitation. NeurIPS 2021.

Additional Work

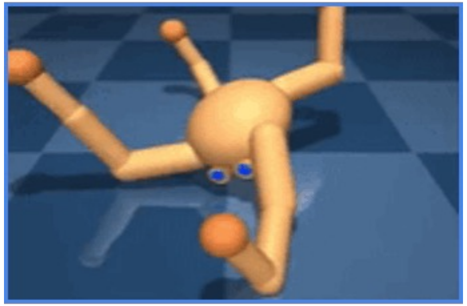
Representation Learning



[1] Nachum and **Yang**. Provable Representation Learning for Imitation. NeurIPS 2021.
[2] **Yang** and Nachum. Offline Pretraining for Sequential Decision Making. ICML 2021.
[3] **Yang** et al. Near-Optimal Imitation with Suboptimal Data. ICLR 2022.
[4] Zhang, Ren, **Yang**, et. al. Linear MDPs via Contrastive Representations. ICML 2022.

Takeaways

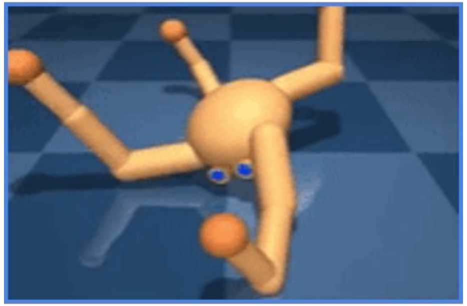
Representation Learning



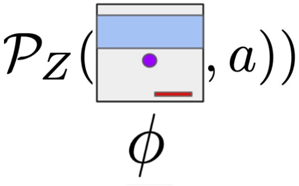
➢ Use suboptimal data for representation learning.

Takeaways

Representation Learning



- ➢ Use suboptimal data for representation learning.
- ➢ Contrastive learning and denoising autoencoding for learning approximate dynamics models.



$$\underline{s, a,}\ s' \Rightarrow s'$$

$$\phi$$

Today's Talk: Foundation Models for Decision Making

Representation Learning

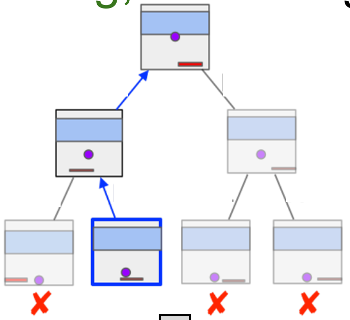
From suboptimal data

[ICML21, NeurIPS21, ICLR22, ICML22]

Reasoning

Input

Planning, search algos



Output

[NeurIPS22]

Internet Data

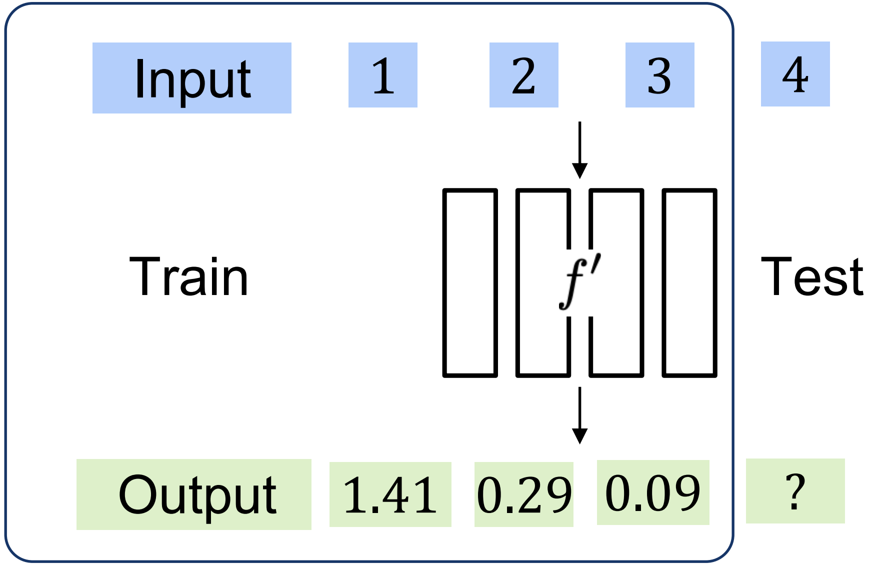
[NeurIPS23, arXiv23, arXiv23]

Teach Models to Do Math

$$f(x) = \frac{x^2 - 1}{x\sqrt{x^2+1}}$$

$f'(x)$?

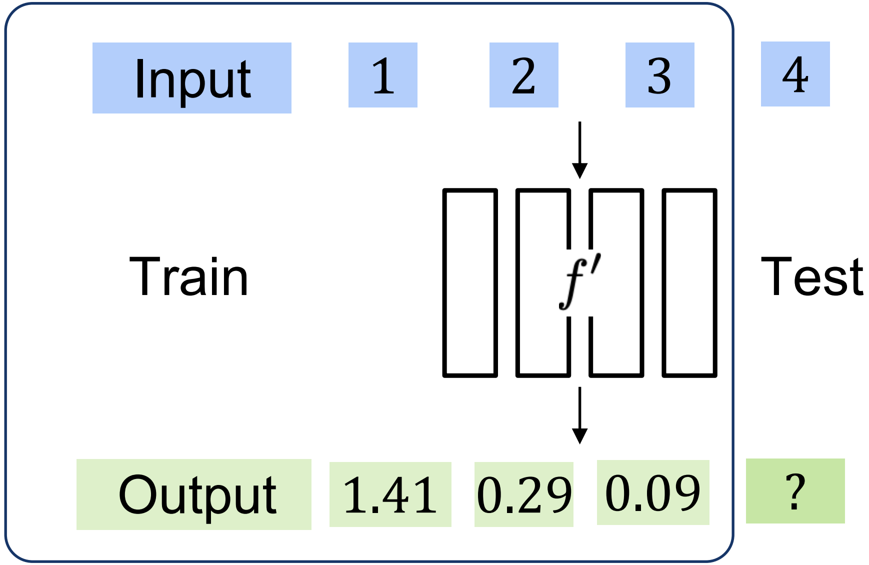
Seems hard!



How Did We Learn Math in School?

$$f(x) = \frac{x^2 - 1}{x\sqrt{x^2 + 1}}$$

$f'(x)$?



How Did We Learn Math in School?

$$f(x) = \frac{x^2 - 1}{x\sqrt{x^2+1}}$$

$f'(x)$?

Quotient rule: $$f'(x) = \frac{(x^2-1)'x\sqrt{x^2+1} - (x^2-1)(x\sqrt{x^2+1})'}{x^2(x^2+1)}$$

Product rule: $$\frac{d}{dx}x\sqrt{x^2+1} = x\frac{d}{dx}\sqrt{x^2+1} + \sqrt{x^2+1}.$$

Chain rule: $$\frac{d}{dx}\sqrt{x^2+1} = \frac{d}{dx}(x^2+1)^{1/2} = \frac{1}{2}(x^2+1)^{-1/2}(2x) = \frac{x}{\sqrt{x^2+1}}.$$

Teach Language Models to Do Math

$$f(x) = \frac{x^2 - 1}{x\sqrt{x^2+1}}$$

4

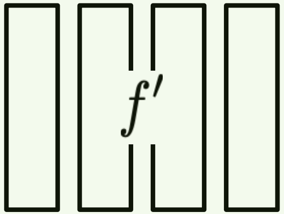
Test

Intermediate reasoning steps

Quotient rule: $$f'(x) = \frac{(x^2-1)'x\sqrt{x^2+1} - (x^2-1)(x\sqrt{x^2+1})'}{x^2(x^2+1)}$$

Product rule: $$\frac{d}{dx}x\sqrt{x^2+1} = x\frac{d}{dx}\sqrt{x^2+1} + \sqrt{x^2+1}.$$

Chain rule: $$\frac{d}{dx}\sqrt{x^2+1} = \frac{d}{dx}(x^2+1)^{1/2} = \frac{1}{2}(x^2+1)^{-1/2}(2x) = \frac{x}{\sqrt{x^2+1}}.$$



$f'(x)$

Understand. Do not memorize.

0.04

[1] Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Language Models. NeurIPS 2022.

How is Math Related to Decision Making?

$$f(x) = \frac{x^2 - 1}{x\sqrt{x^2+1}}$$

4

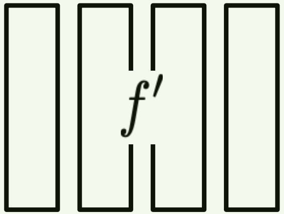
Test

Intermediate reasoning steps

Quotient rule: $$f'(x) = \frac{(x^2-1)'x\sqrt{x^2+1} - (x^2-1)(x\sqrt{x^2+1})'}{x^2(x^2+1)}$$

Product rule: $$\frac{d}{dx}x\sqrt{x^2+1} = x\frac{d}{dx}\sqrt{x^2+1} + \sqrt{x^2+1}.$$

Chain rule: $$\frac{d}{dx}\sqrt{x^2+1} = \frac{d}{dx}(x^2+1)^{1/2} = \frac{1}{2}(x^2+1)^{-1/2}(2x) = \frac{x}{\sqrt{x^2+1}}.$$



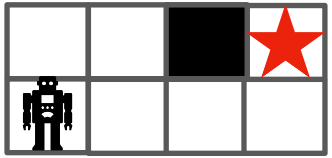
$f'(x)$

Understand. Do not memorize.

0.04

[1] Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Language Models. NeurIPS 2022.

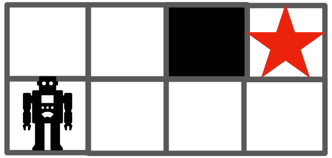
Teach Models to Search

s 

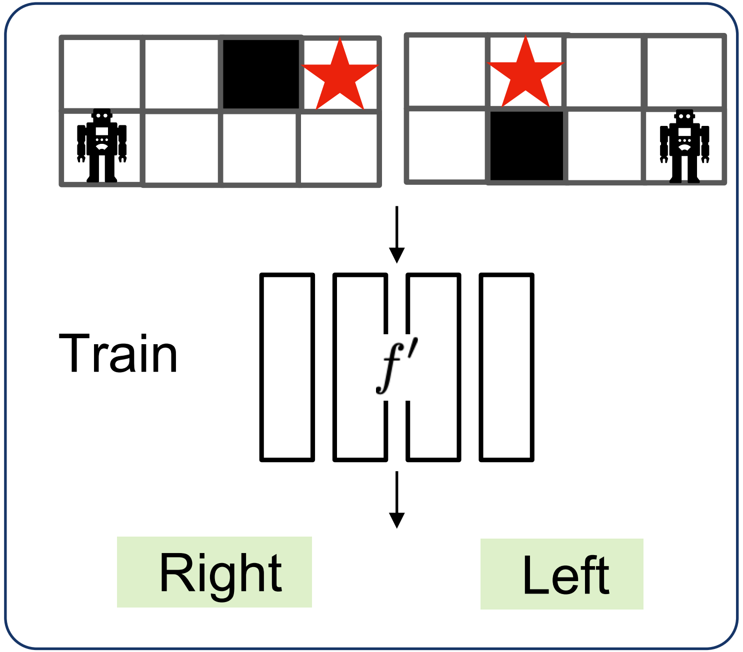
$a\,?$

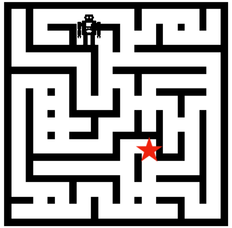
Teach Models to Search via Behavioral Cloning

s



$a\,?$





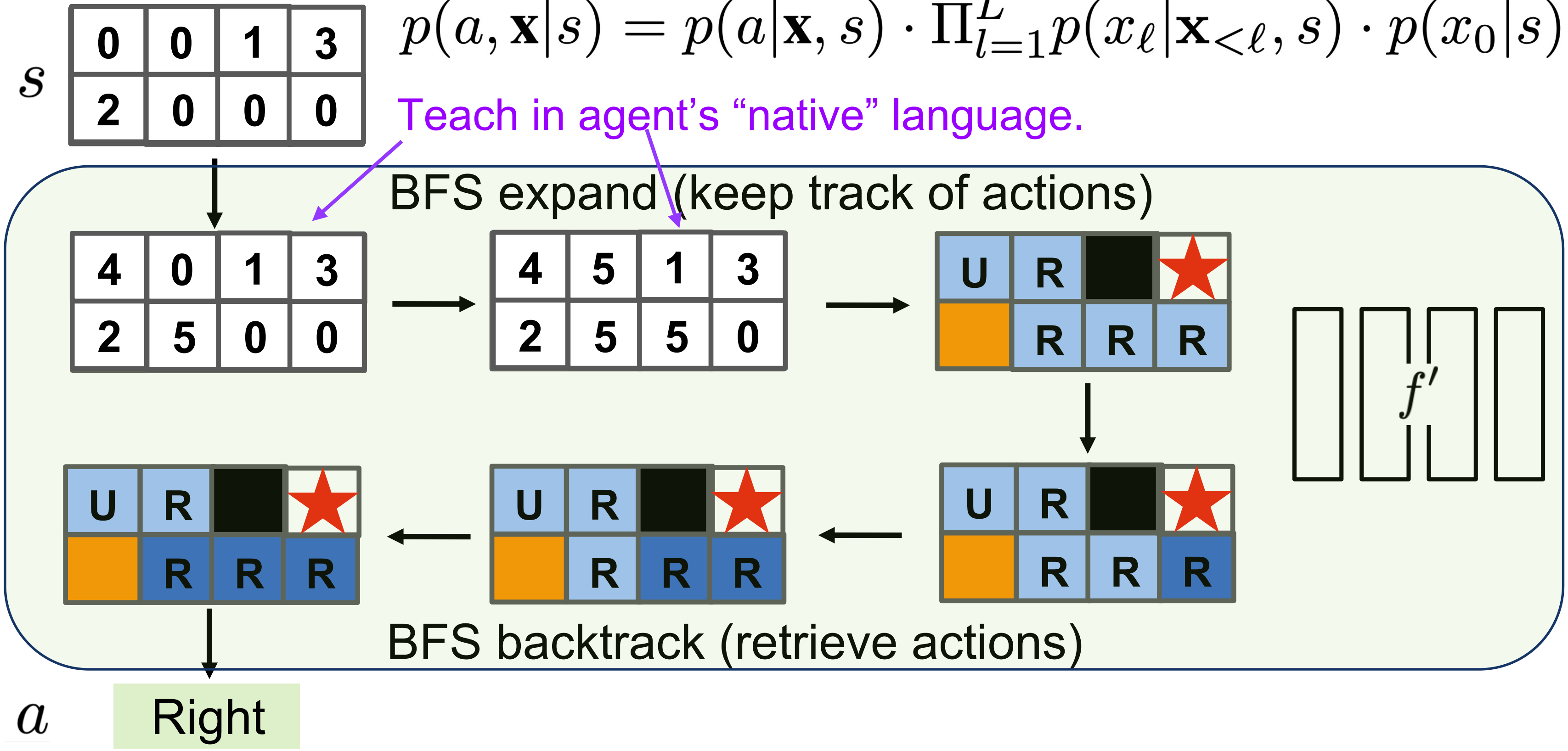
Test

?

Teach Models to Search via Procedure Cloning

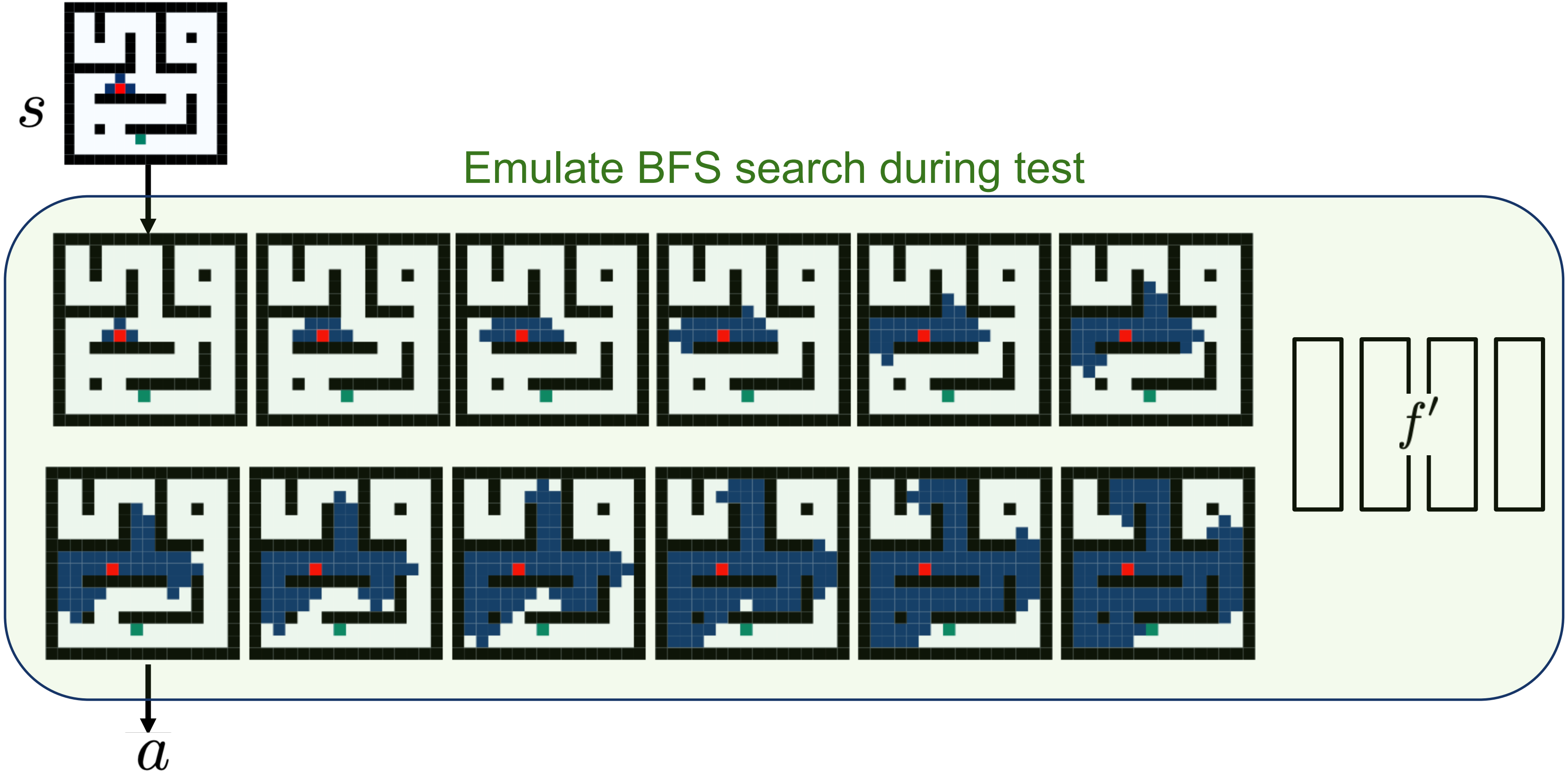
$$p(a, \mathbf{x}|s) = p(a|\mathbf{x}, s) \cdot \Pi_{l=1}^{L} p(x_\ell|\mathbf{x}_{<\ell}, s) \cdot p(x_0|s)$$

Teach in agent's "native" language.



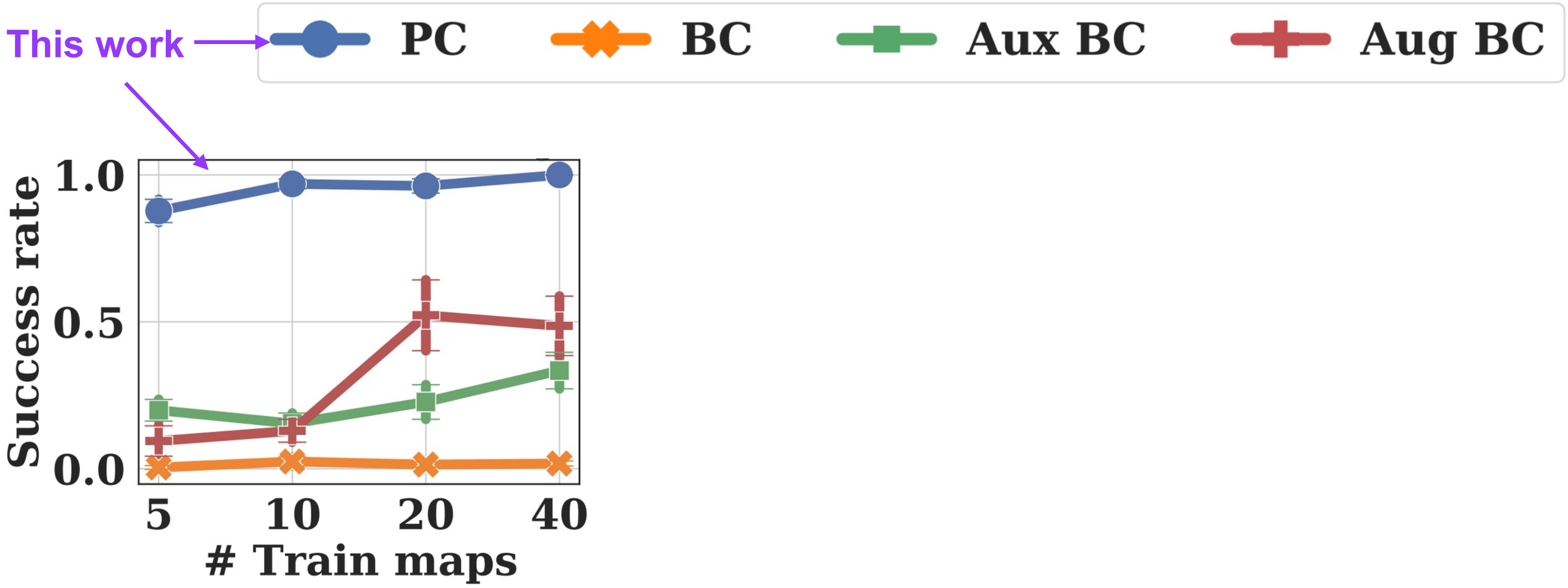
[1] **Yang** et al. Chain-of-Thought Imitation with Procedure Cloning. NeurIPS 2022.

Teach Models to Search via Procedure Cloning



[1] **Yang** et al. Chain-of-Thought Imitation with Procedure Cloning. NeurIPS 2022.

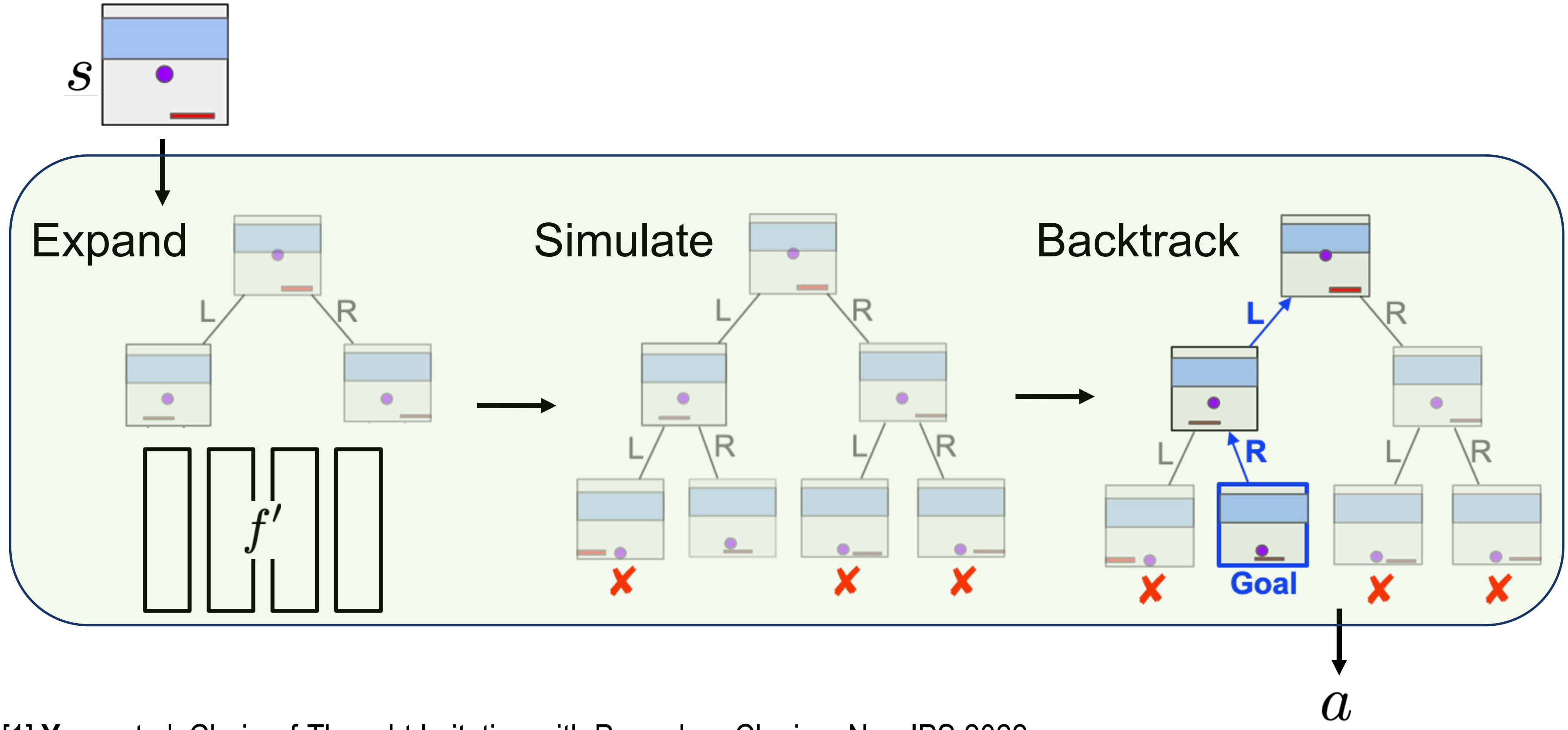
Empirical Performance of Procedure Cloning



Navigation

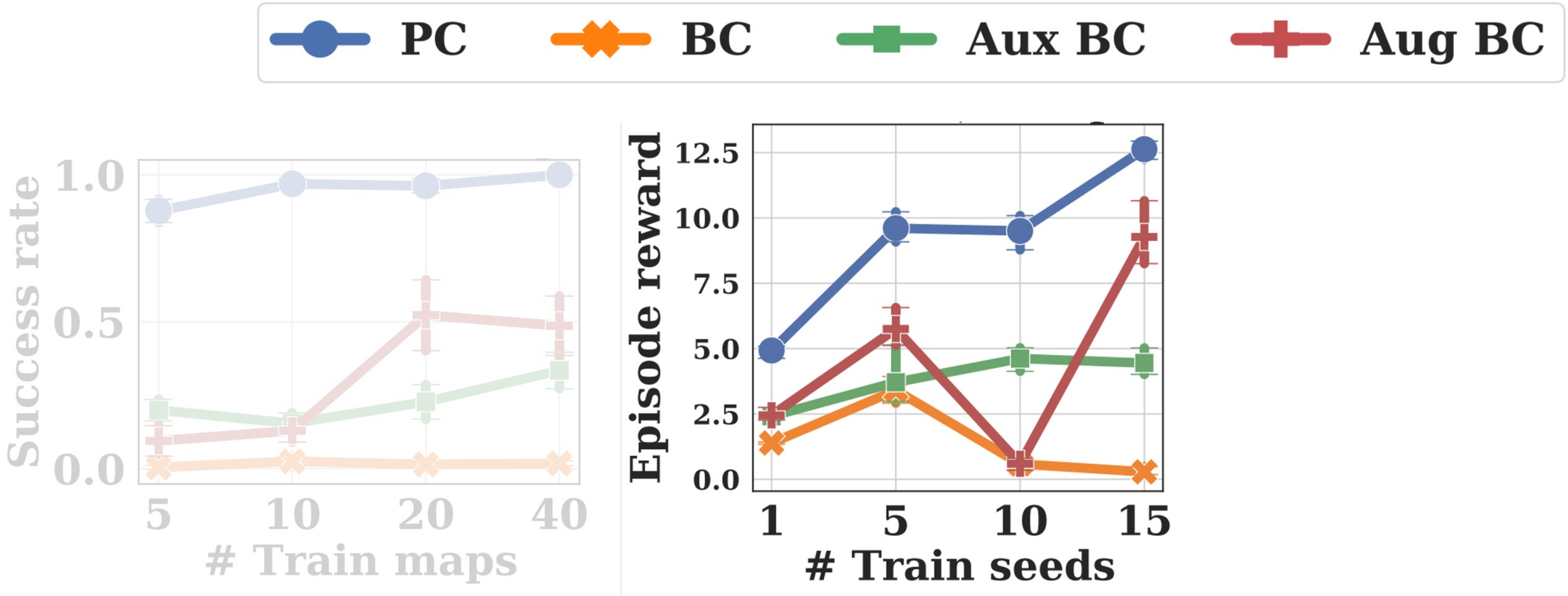
[1] **Yang** et al. Chain-of-Thought Imitation with Procedure Cloning. NeurIPS 2022.

Procedure Cloning is General: MCTS



[1] **Yang** et al. Chain-of-Thought Imitation with Procedure Cloning. NeurIPS 2022.

Empirical Performance of Procedure Cloning

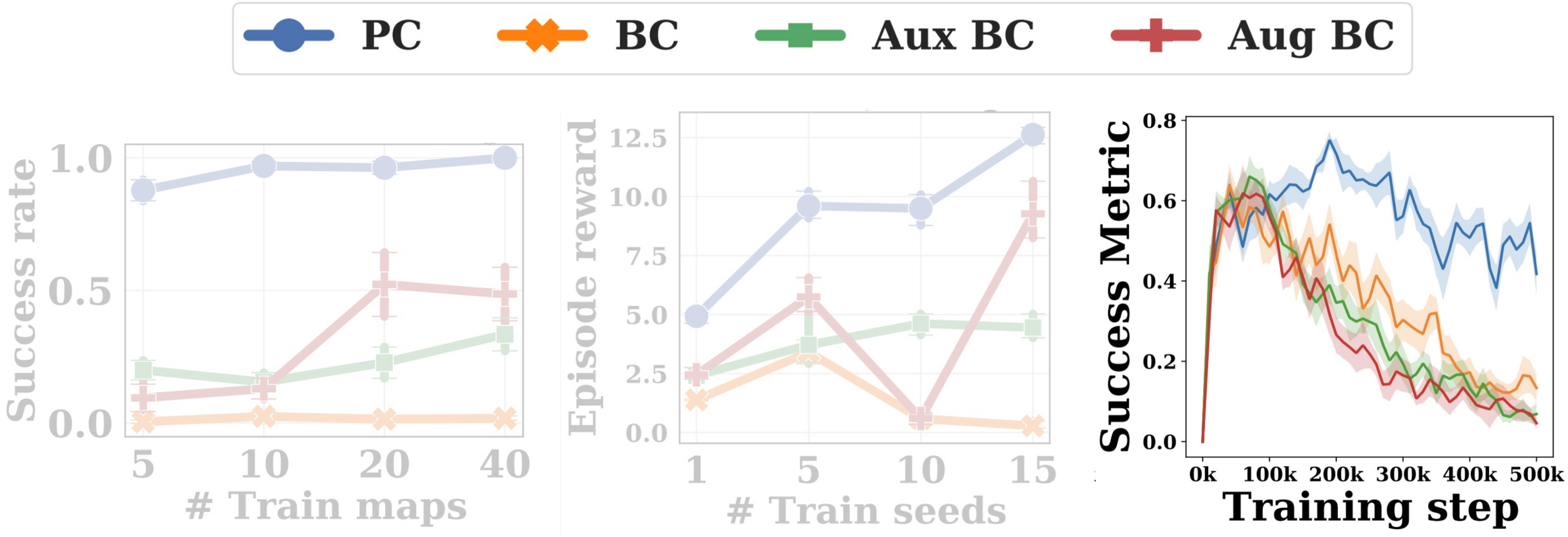


Navigation

Games

[1] **Yang** et al. Chain-of-Thought Imitation with Procedure Cloning. NeurIPS 2022.

Empirical Performance of Procedure Cloning



Navigation

Games

Manipulation

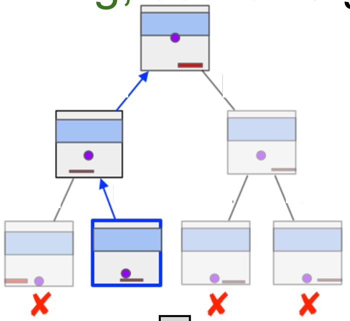
[1] **Yang** et al. Chain-of-Thought Imitation with Procedure Cloning. NeurIPS 2022.

Takeaways

Reasoning in Agents

State

Planning, search algos



Action

➢ Teach intermediate computations.

Takeaways

Reasoning in LLMs

Input

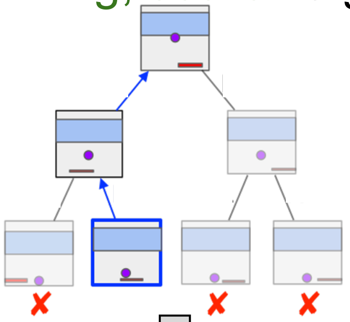
Natural language steps

Output

Reasoning in Agents

State

Planning, search algos



Action

- ➢ Teach intermediate computations.
- ➢ Don’t need to teach in human language. Teach in machine language.

Today's Talk: Foundation Models for Decision Making

Representation Learning

From suboptimal data

[ICML21, NeurIPS21, ICLR22, ICML22]

Reasoning

Input

Planning, search algos

Output

[NeurIPS22]

Internet Data

[NeurIPS23, arXiv23, arXiv23]

Human-Like Chatbot from Internet Language Data

Internet language data

WIKIPEDIA

Common Crawl

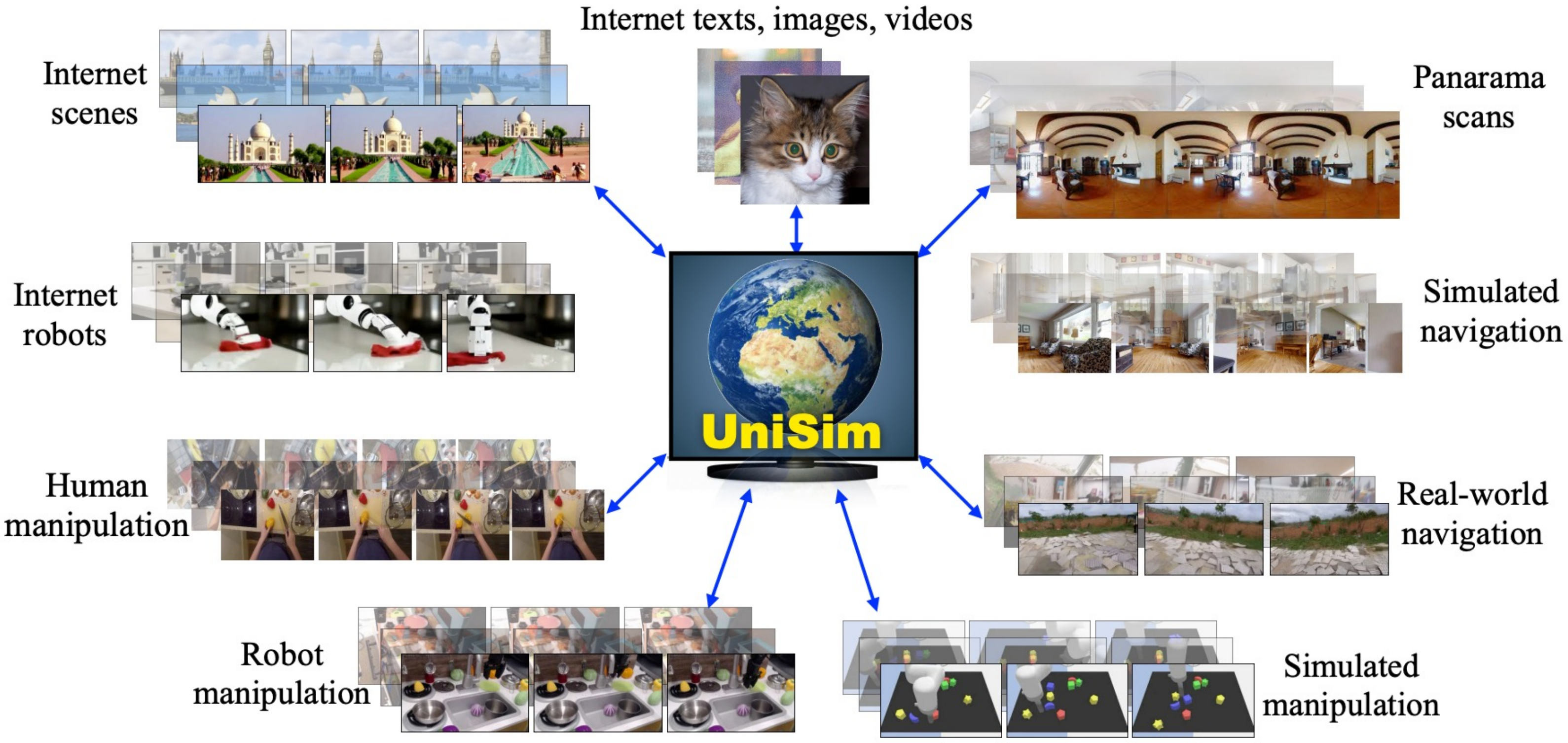
Human-like chatbot

JI What are Foundation Models?

Foundation models are large pre-trained neural networks used in machine learning and natural language processing. They form the foundation for various tasks and are trained on extensive internet text data, enabling them to grasp a wide range of knowledge and language patterns. Prominent examples include OpenAI's GPT series and Google's BERT model.

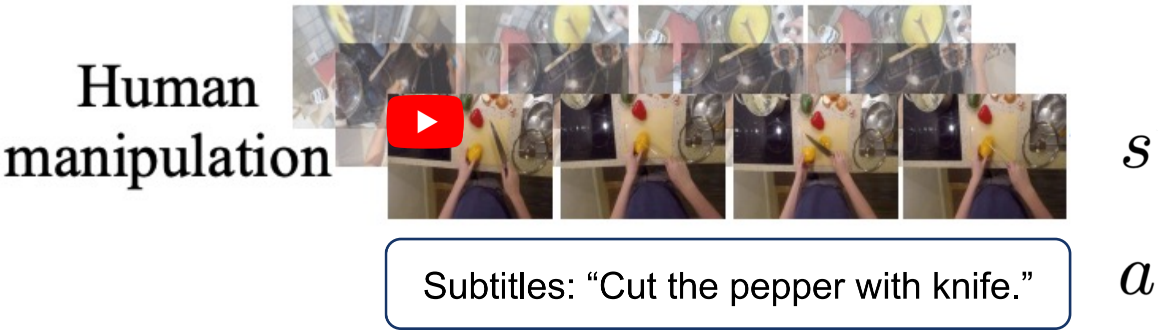
World-Like Simulator from Internet Multimodal Data?

Different state action spaces.



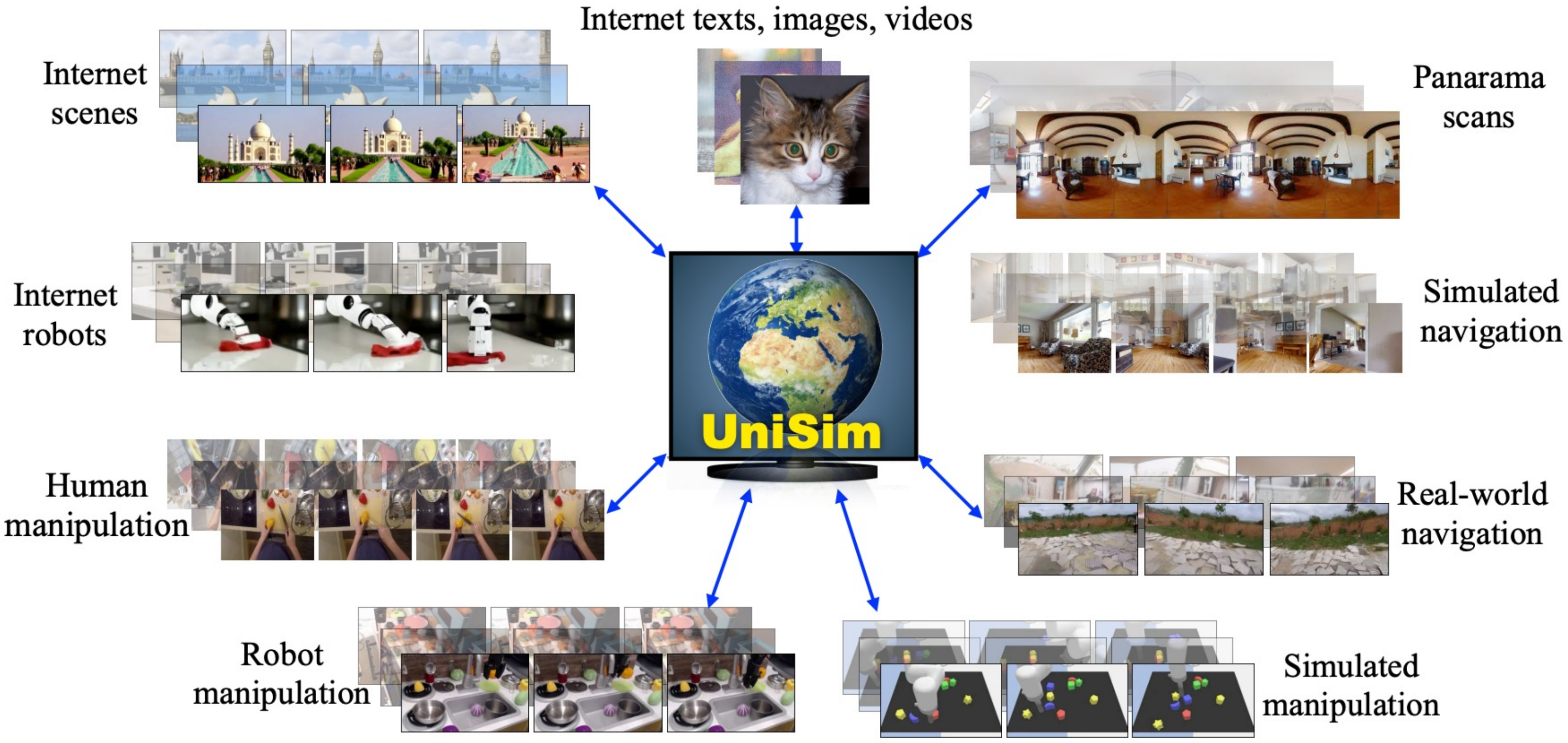
[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Video and Text as Universal State and Action



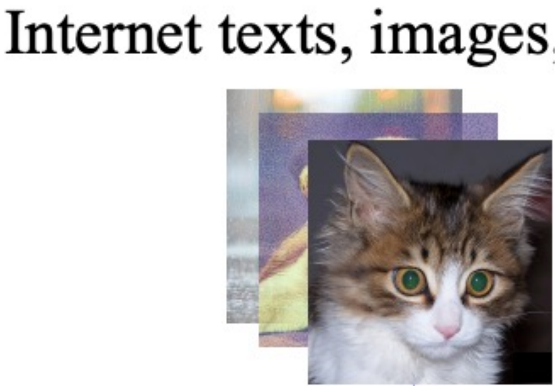
[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Video and Text as Universal State and Action



[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Video and Text as Universal State and Action



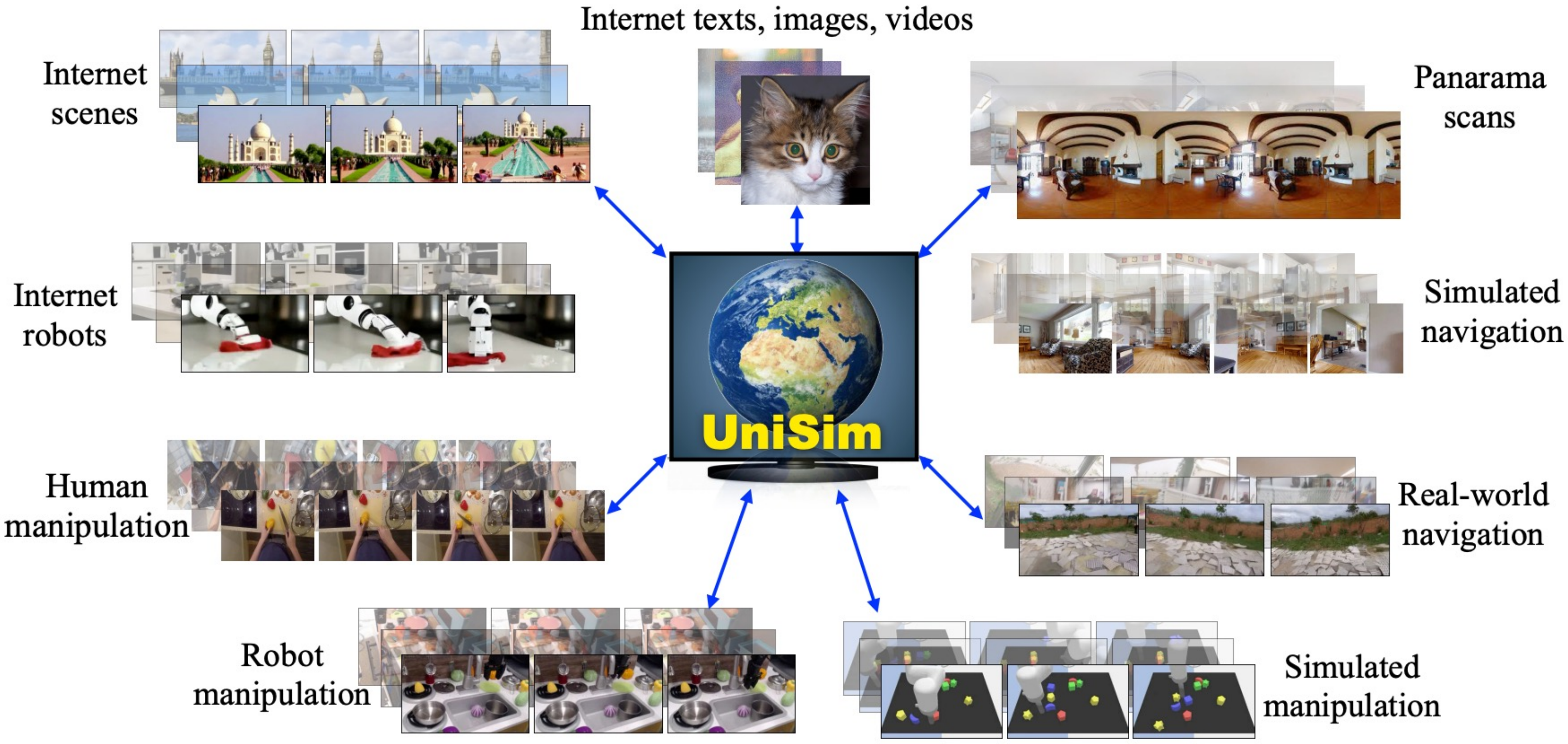
s

Caption: “A cat staring straight.”

a

[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Video and Text as Universal State and Action



[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Video and Text as Universal State and Action

Panarama scans

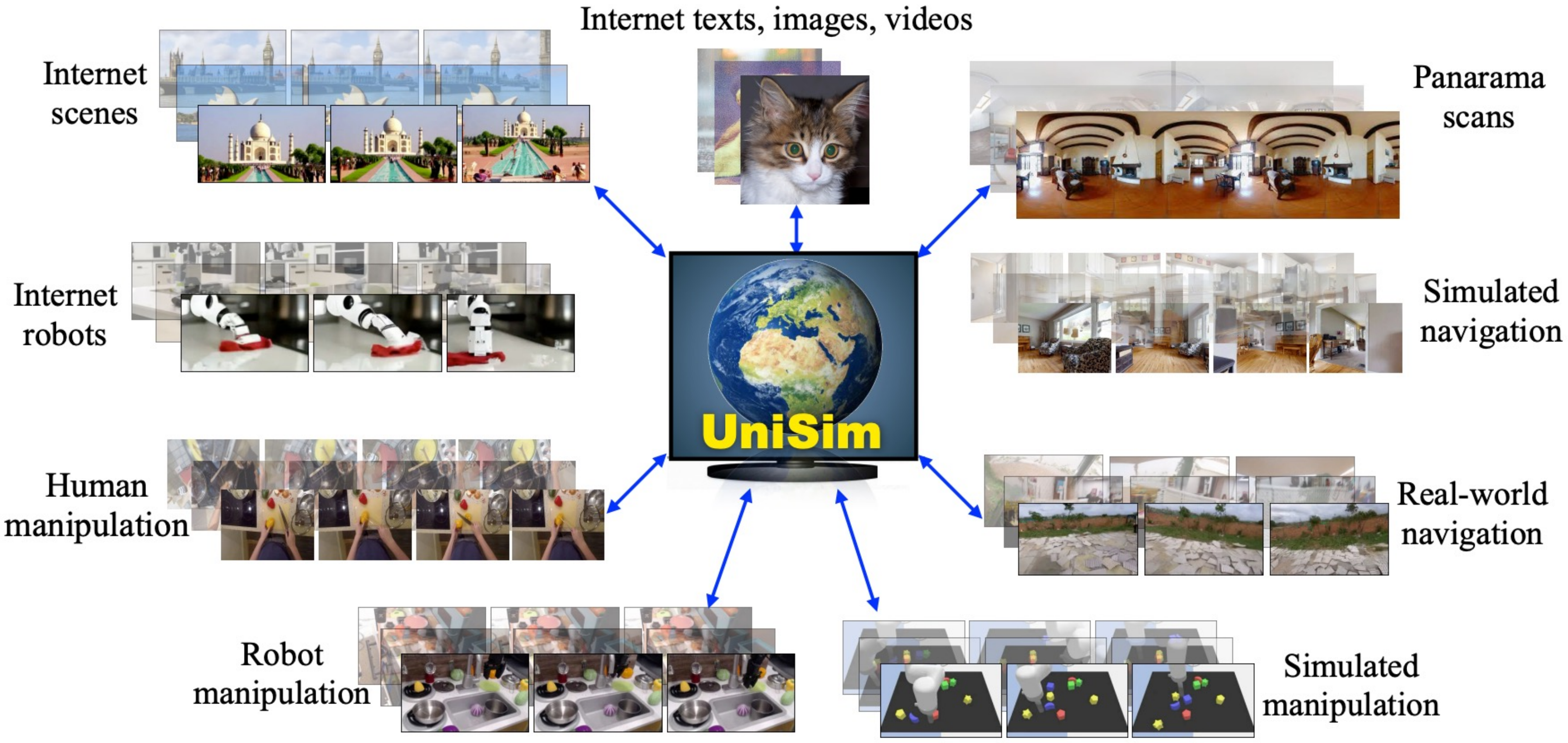
s

<camera> 90°, <zoom> 1.5

a

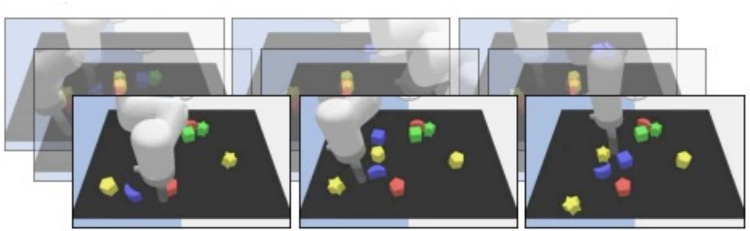
[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Video and Text as Universal State and Action



[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Video and Text as Universal State and Action



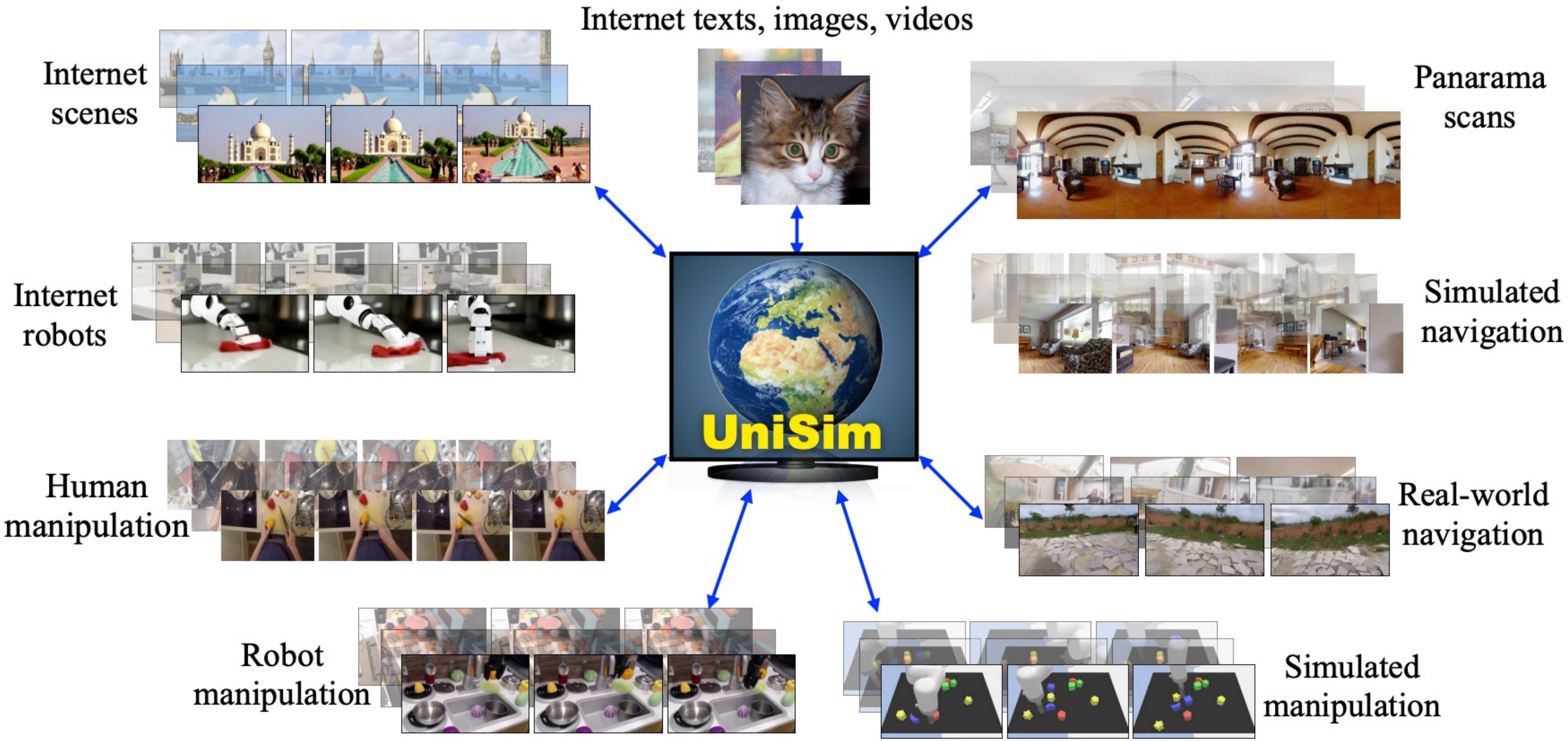
s

$\Delta x, \Delta y$

a

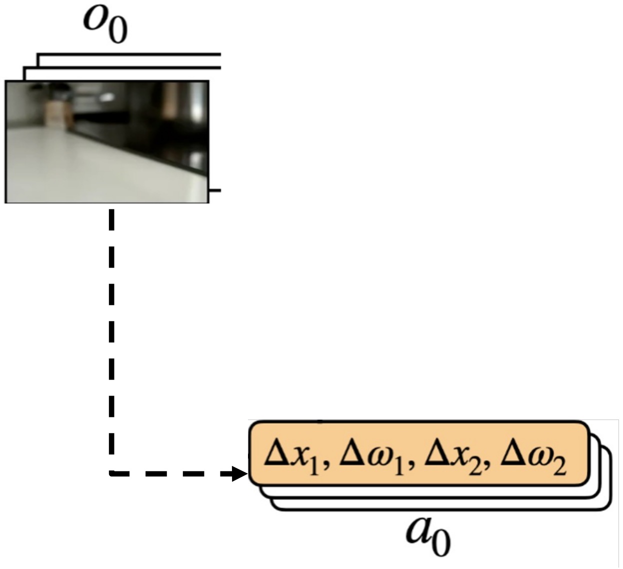
[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Video and Text as Universal State and Action



[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Text-to-Video Generation as a Universal Simulator



[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

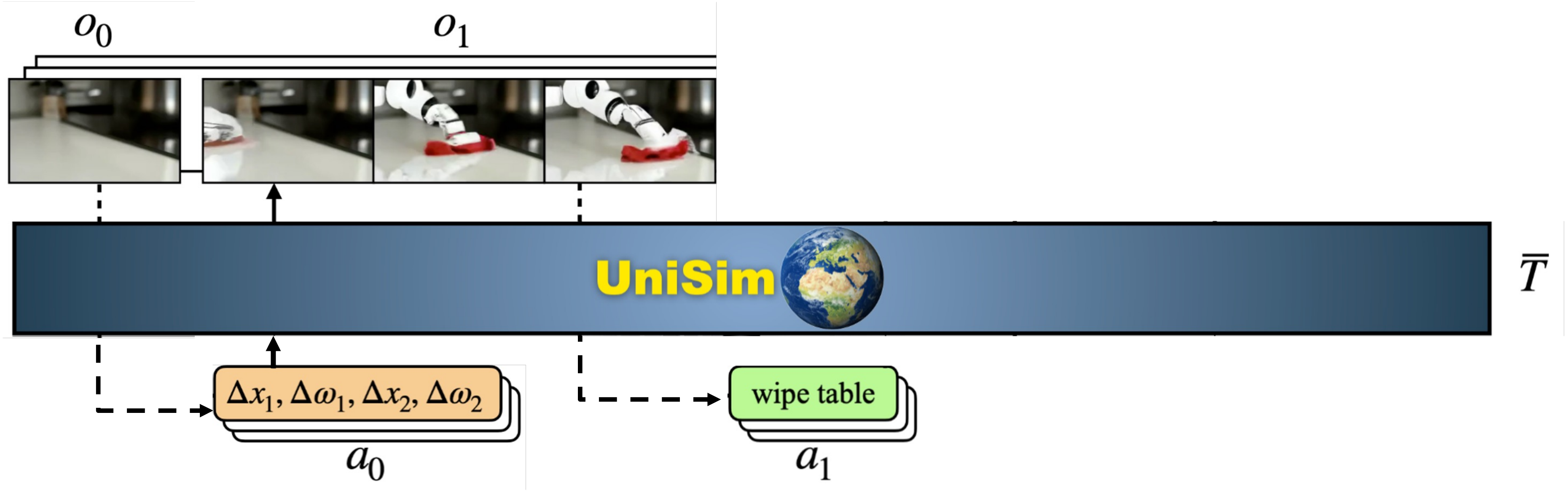
Text-to-Video Generation as a Universal Simulator

[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Text-to-Video Generation as a Universal Simulator

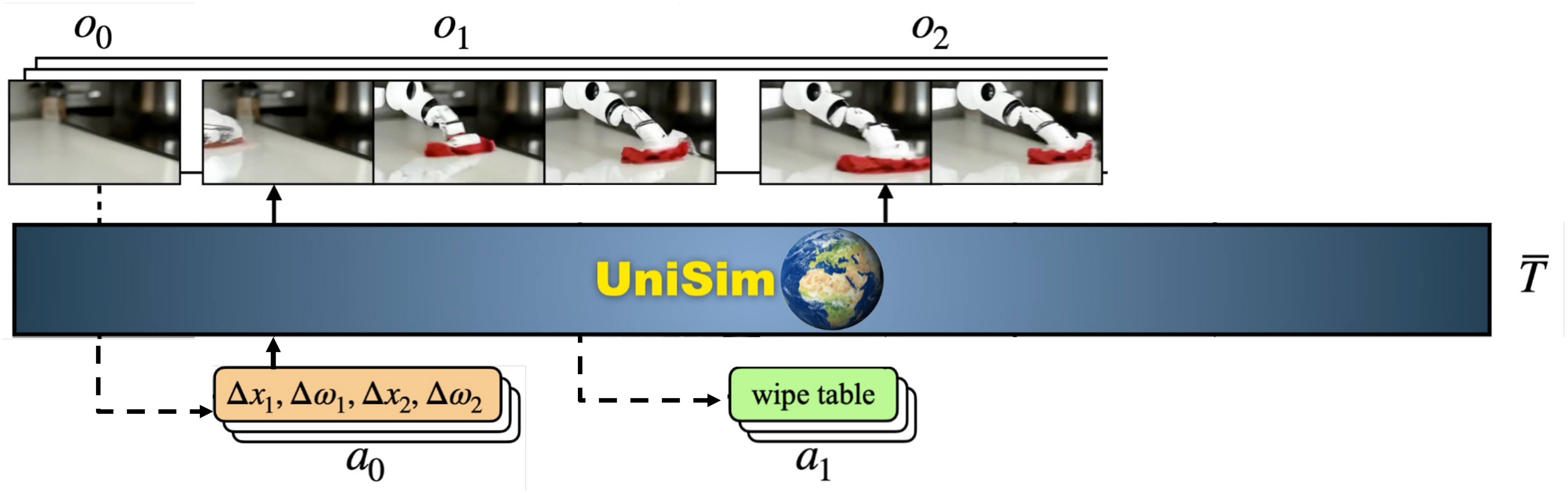
[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Text-to-Video Generation as a Universal Simulator



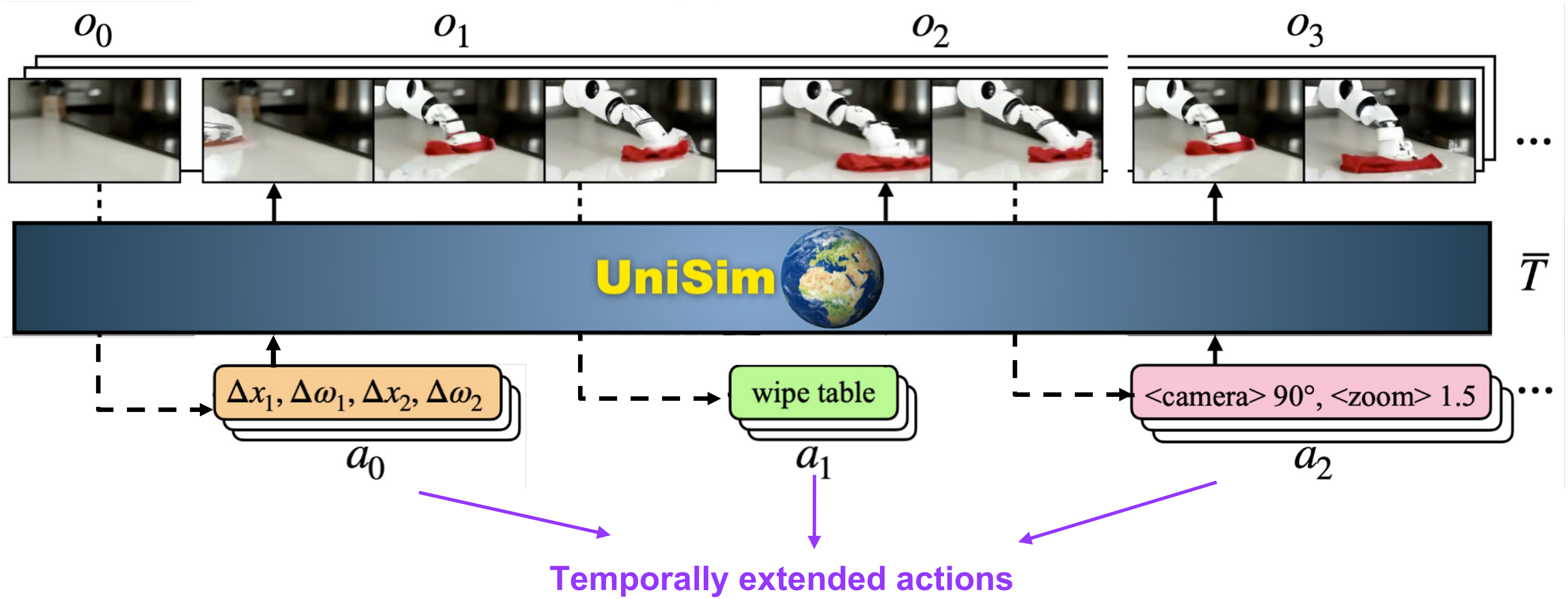
[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Text-to-Video Generation as a Universal Simulator



[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Text-to-Video Generation as a Universal Simulator



[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

UniSim Demos

[Demo Link]

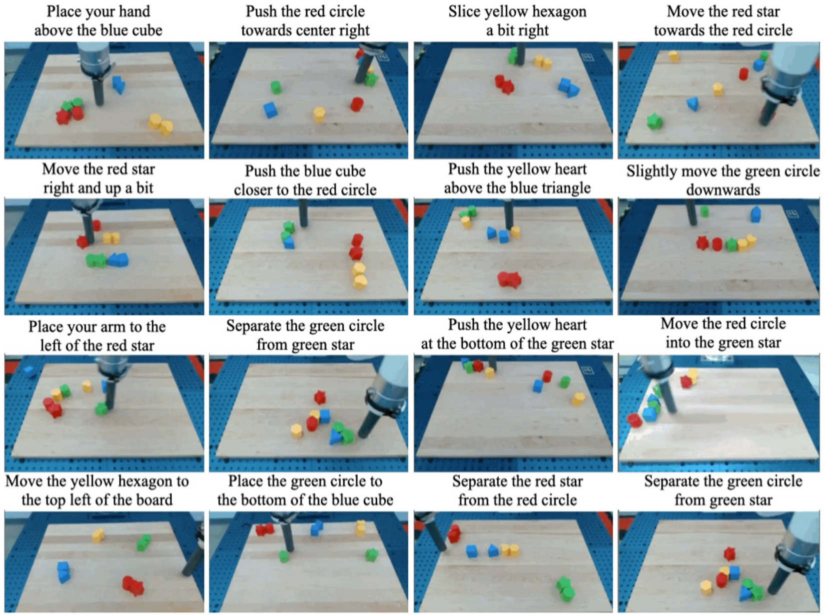
[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Application: Large-Scale “Online” RL

Challenge: Sample Efficiency

Challenge: Generalization

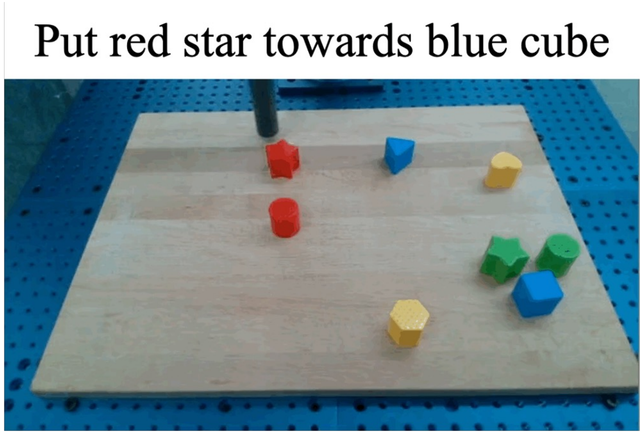
Universal Simulator



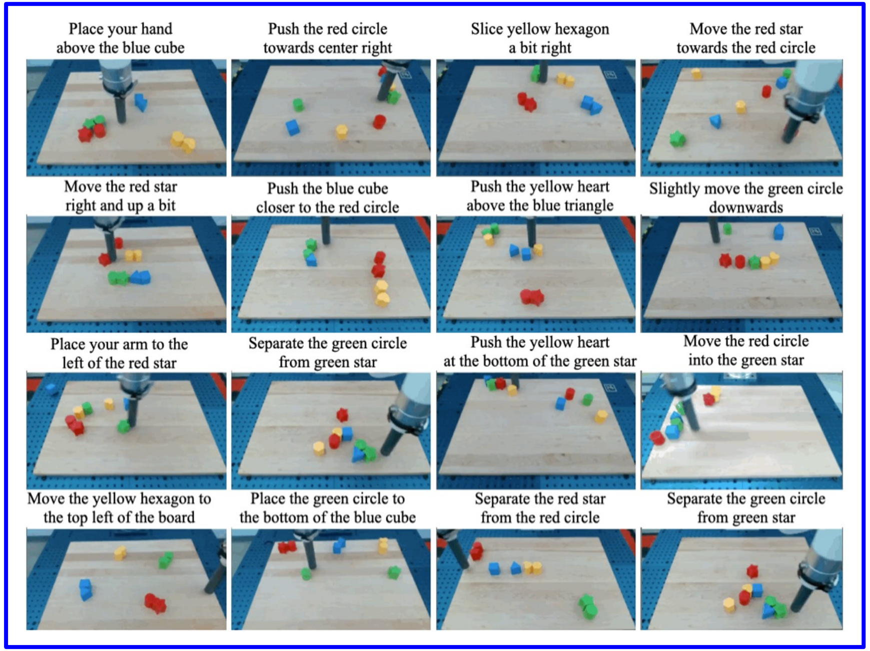
[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Application: Large-Scale “Online” RL

Zero-shot real-world transfer



Universal Simulator



[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.

Application: Search and Planning

Put the fruits into the top drawer

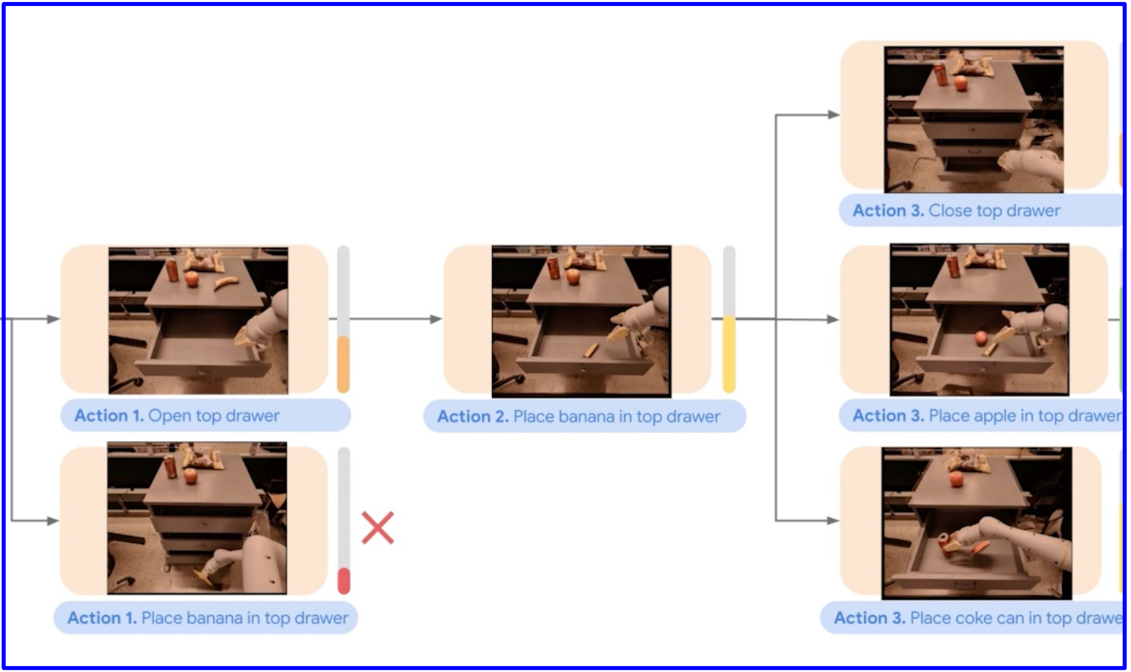
Search and planning in simulation

[1] Du*, **Yang*** et al. Learning Universal Policies via Text-Guided Video Generation. NeurIPS 2023.
[2] Du, **Yang**, et al. Video Language Planning. arXiv2023.

Application: Search and Planning

Zero-shot real-world transfer

Search and planning in simulation



[1] Du*, **Yang*** et al. Learning Universal Policies via Text-Guided Video Generation. NeurIPS 2023.
[2] Du, **Yang**, et al. Video Language Planning. arXiv2023.

Takeaways

Internet Data

➢ Rich interactive data on the internet to improve decision making.

Takeaways

Internet Data

- ➢ Rich interactive data on the internet to improve decision making.
- ➢ LLMs, VLMs, text-to-video models parametrize different components of MDPs.

Takeaways

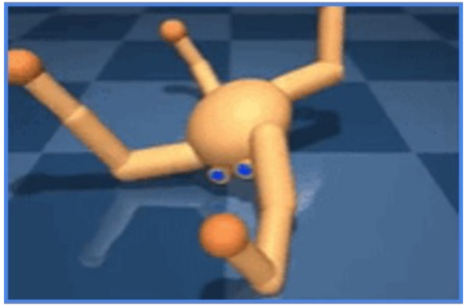
Internet Data

- ➢ Rich interactive data on the internet to improve decision making.
- ➢ LLMs, VLMs, text-to-video models parametrize different components of MDPs.
- ➢ Large-scale “online” access through generative modeling for RL, search, planning.

Foundation Models for Control and Embodiment

Representation Learning

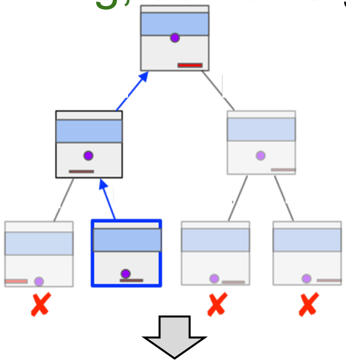
From suboptimal data



Reasoning

Input

Planning, search algos



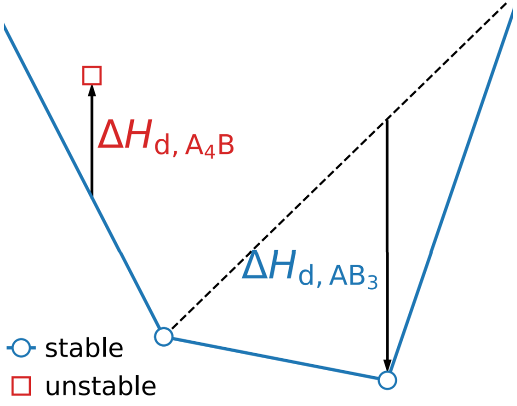
Output

Internet Data

Foundation Models for Materials Discovery

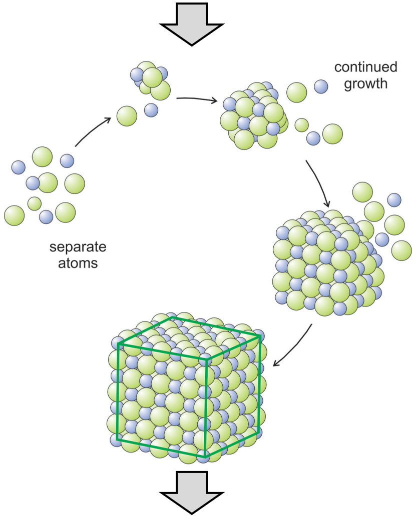
Representation Learning

From suboptimal data



Reasoning

Input



Output

Internet Data

ICSD

OQMD
The Open Quantum Materials Database

The Materials Project

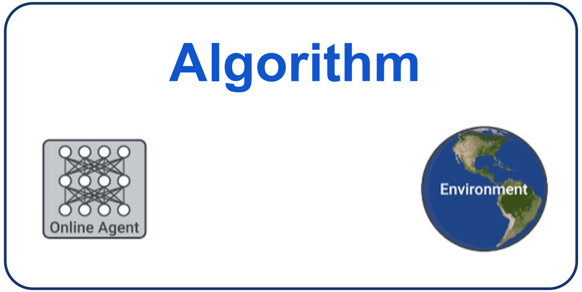
[1] **Yang** et al. Scalable Diffusion for Materials Discovery. arXiv 2023.

Big Picture: The Past and Future of FMDM

Algorithm: RL, planning, control, optimization.

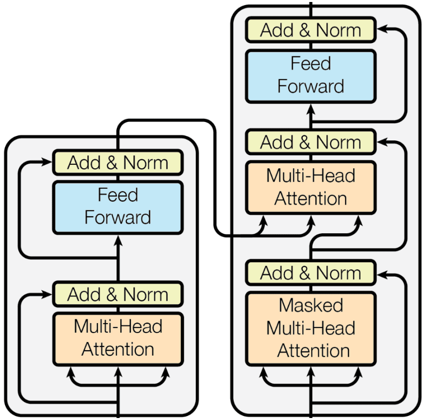
[1] **Yang***, Nachum*, Dai* et al. Off-Policy Evaluation via the Regularized Lagrangian. NeurIPS 2020.
[2] **Yang***, Dai*, Nachum* et al. Offline Policy Selection under Uncertainty. AISTATS 2022.

Big Picture: The Past and Future of FMDM



Big Picture: The Past and Future of FMDM

Model: Attention, transformers, autoregressive, diffusion.



Transformer agent

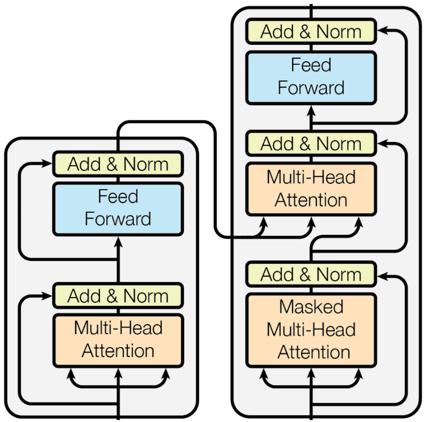
Multi-task environments

[1] Lee*, Nachum*, **Yang** et al. Multi-Game Decision Transformers. NeurIPS 2022.
[2] **Yang** et al. Dichotomy of Control. ICLR 2023.
[3] Venuto*, **Yang***, et al. Multi-Environment Pretraining Enables Transfer to Action Limited Datasets

Big Picture: The Past and Future of FMDM

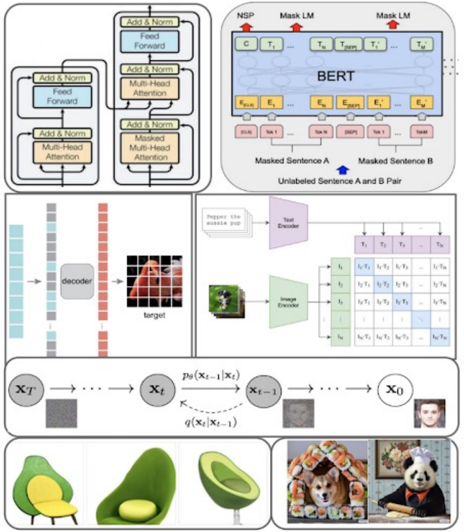
Algorithm

Model



Big Picture: The Past and Future of FMDM

Data: Internet text, image, video, action.

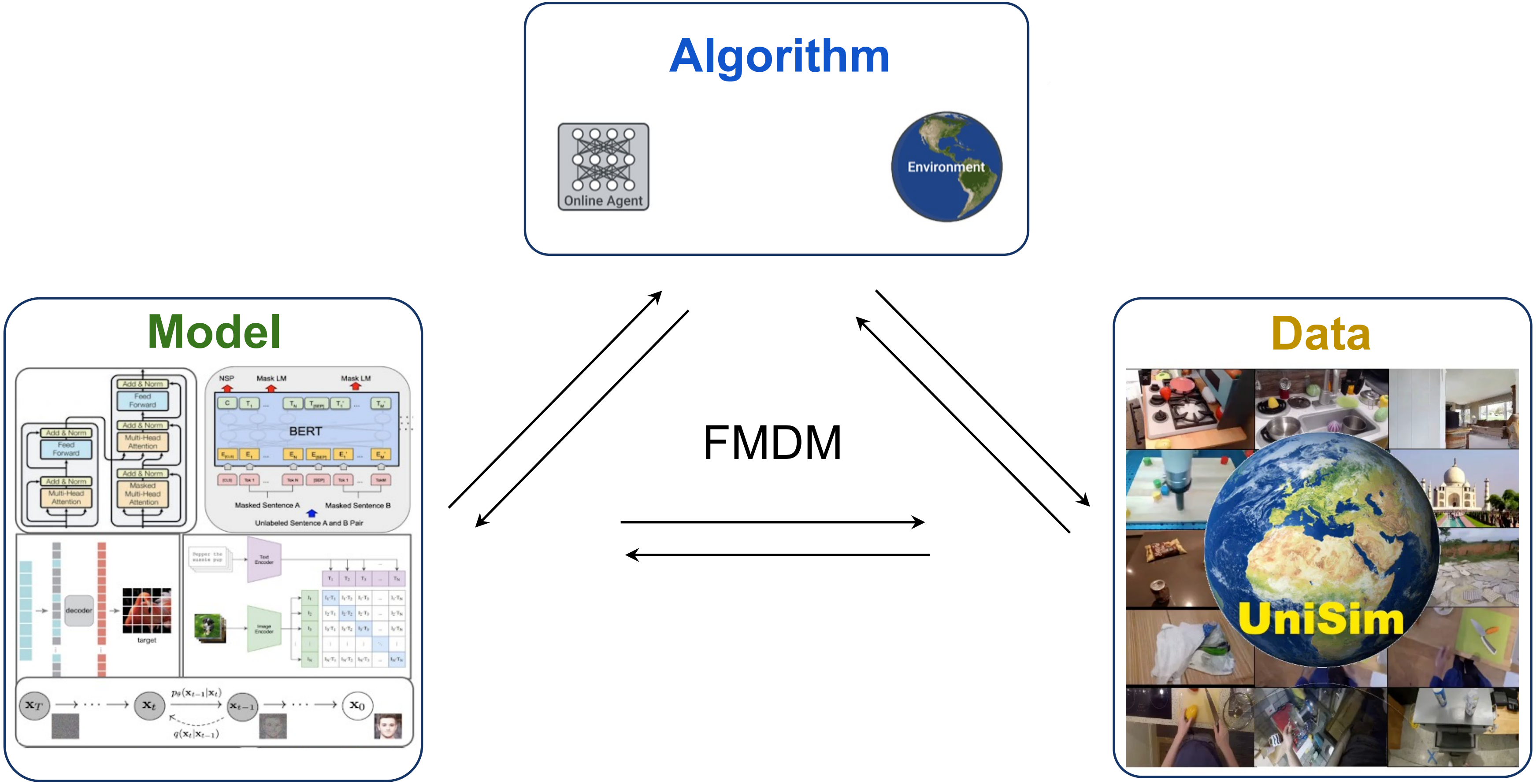


Foundation agent model

Foundation world model

[1] **Yang** et al. Learning Interactive Real-World Simulators. arXiv 2023.
[2] Du*, **Yang*** et al. Learning Universal Policies via Text-Guided Video Generation. NeurIPS 2023.
[3] Du, **Yang**, et al. Video Language Planning. arXiv2023.

Big Picture: The Past and Future of FMDM

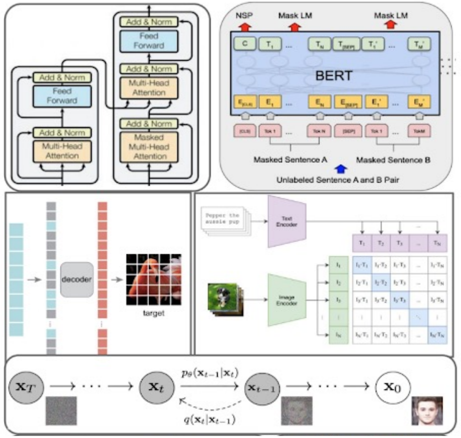


Big Picture: The Past and Future of FMDM

Algorithm

- Algorithm guarantees relies on assumptions of modelling flexibility and data coverage.

Model

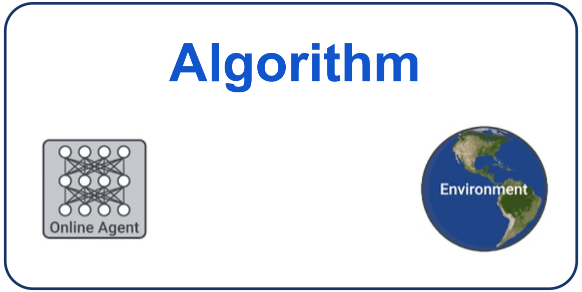


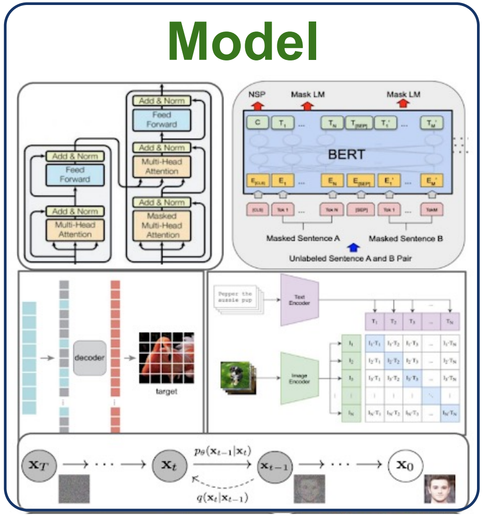
FMDM

Data

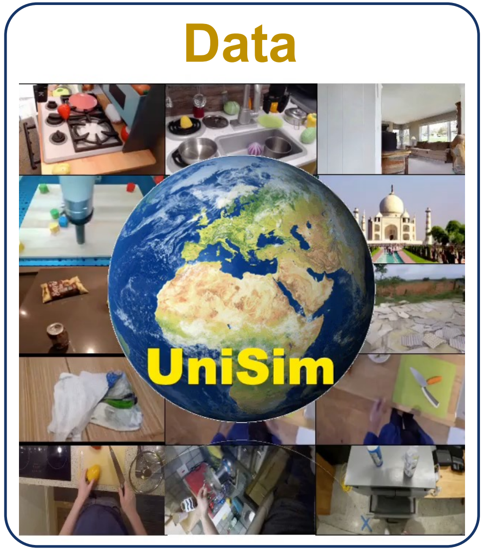
Big Picture: The Past and Future of FMDM

- Models are improved by algorithms (RLHF) and interactive data.





FMDM

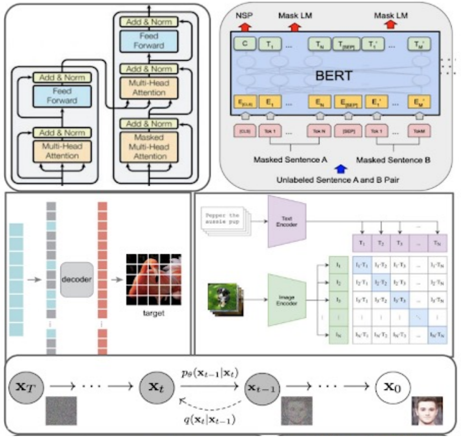


[1] **Yang***, Du* et al. Probabilistic Adaptation of Text-to-Video Models. arXiv 2023.

Big Picture: The Past and Future of FMDM

Algorithm

Model

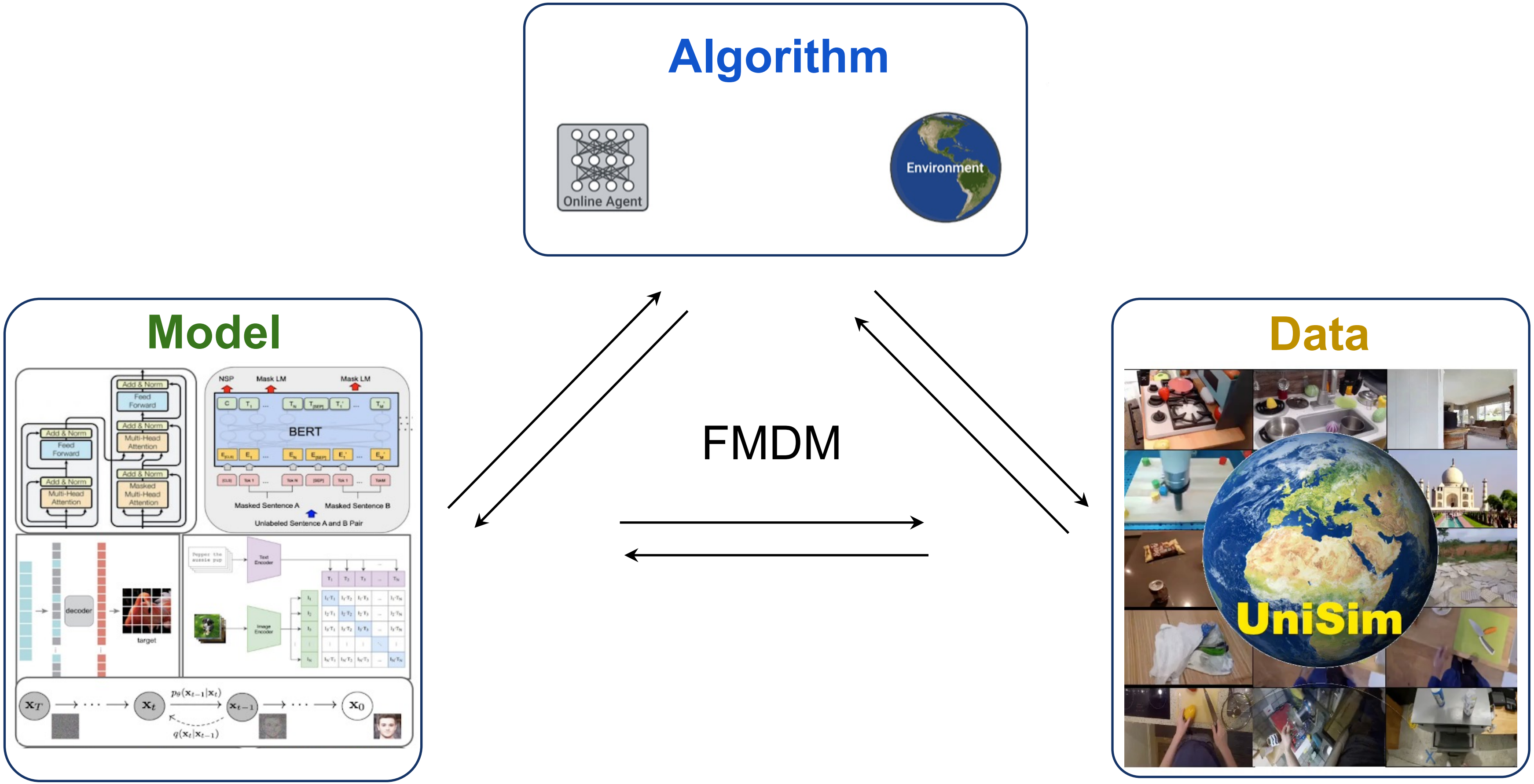


FMDM

Data

- New data are produced / generated by deploying models and running algorithms.

Thank You


Algorithm
Online Agent
Environment
FMDM
Model
Add & Norm
Feed Forward
Multi-Head Attention
Masked Multi-Head Attention
NSP
Mask LM
BERT
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
decoder
target
Data
UniSim

Summary

Representation Learning

➢ Learn dynamics and state representations.

Reasoning

➢ Learn intermediate steps of algorithms.

Internet Data

➢ Learn large-scale agents and simulators.

Outlook

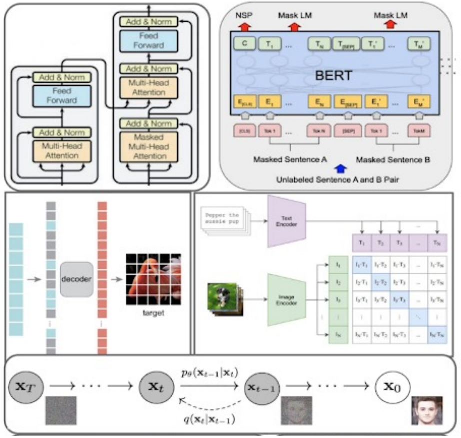
Algorithm

➢ RL, search, planning.

Model

➢ MLPs, RNNs.

➢ Transformers, foundation models



Data

➢ (Single) task-specific.

➢ Multi-task, internet

Technical Details

- Representation learning
    - [Sample efficiency]
    - [Contrastive learning and random Fourier features]

Representation Learning Sample Efficiency

Lemma 11. *Let $\rho \in \Delta(\{1, \ldots, k\})$ be a distribution with finite support. Let $\hat{\rho}_n$ denote the empirical estimate of ρ from n i.i.d. samples $X \sim \rho$. Then,*

$$\mathbb{E}_n[D_{\mathrm{TV}}(\rho \| \hat{\rho}_n)] \leq \frac{1}{2} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \sqrt{\rho(i)} \leq \frac{1}{2} \cdot \sqrt{\frac{k}{n}}. \tag{66}$$

Lemma 12. *Let $\mathcal{D} := \{(s_i, a_i)\}_{i=1}^n$ be i.i.d. samples from a factored distribution $x(s, a) := \rho(s)\pi(a|s)$ for $\rho \in \Delta(S), \pi : S \to \Delta(A)$. Let $\hat{\rho}$ be the empirical estimate of ρ in $\mathcal{D}$ and $\hat{\pi}$ be the empirical estimate of π in $\mathcal{D}$. Then,*

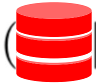
$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{s\sim\rho}[D_{\mathrm{TV}}(\pi(s) \| \hat{\pi}(s))]] \leq \sqrt{\frac{|S||A|}{n}}. \tag{67}$$

Theorem 4. *Consider the setting described above. Let $\phi_M := \mathcal{OPT}_\phi(\mathcal{D}_M^{\mathrm{off}})$ and $\pi_{N,Z}$ be the policy resulting from BC with respect to ϕ_M. Then we have,*

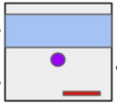
$$\mathbb{E}_{\mathcal{D}^{\pi_*}}[\mathrm{PerfDiff}(\pi_{N,Z}, \pi_*)] \leq (1 + D_{\chi^2}(d^{\mathrm{off}} \| d^{\pi_*})^{\frac{1}{2}}) \cdot \epsilon_{\mathrm{R,T}}(\phi_M) + \boxed{C \cdot \sqrt{\frac{|Z||A|}{N}},} \tag{15}$$

where $C = \frac{2R_{\max}}{(1-\gamma)^2}$

Contrastive Learning and Random Fourier Features

$\mathrm{PerfDiff}(\pi_Z, \pi_*) \leq (1 + D_{\chi^2}($  $\|$  $)^{\frac{1}{2}}) \cdot \boxed{\epsilon_{\mathrm{R,T}}} + C\sqrt{\frac{1}{2}\underbrace{\mathbb{E}_{z\sim d_Z^{\pi_*}}[D_{\mathrm{KL}}(\pi_{*,Z}(z)\|\pi_Z(z))]}_{= \ \mathrm{const}(\pi_*, \phi) + J_{\mathrm{BC},\phi}(\pi_Z)}}$

Approx. dynamics model

$D_{\mathrm{KL}}(\mathcal{P}($  $, a)\|\mathcal{P}_Z($  $, a))$

$$\overline{P}(s'|s,a) \propto \rho(s')\exp\{-||\phi(s) - g(s',a)||^2\}$$

Define approx. dynamics model as EBM.

$$\Downarrow$$

Minimizing KL reduces to contrastive learning.

$$D_{\mathrm{KL}}(P(s,a)\|\overline{P}(s,a)) = \mathbb{E}_{s'\sim P(s,a)}[||\phi(s) - g(s',a)||^2] + \log \mathbb{E}_{\tilde{s}'\sim\rho}\exp\{-||\phi(s) - g(\tilde{s}',a)||^2\}$$

$$\overline{P}(s'|s,a) \propto \rho(s')\exp\{-||\phi(s) - g(s',a)||^2\} \boxed{\approx \rho(s')\cdot\varphi(\phi(s))^\top\varphi(g(s',a))}$$

Recover linearization via random Fourier features.

Contrastive Learning and Random Fourier Features

Theorem: For any target policy $\boldsymbol{\pi}^*$, representation ϕ, policy $\pi_\theta(z) := \mathrm{softmax}(\theta^\top z)$ and model error $\epsilon_{\mathrm{R,T}}$ measured with linear dynamics models:

$$\underbrace{\boxed{\mathrm{PerfDiff}(\pi_\theta, \pi_*)}}_{\text{Learning Goal}} \le \underbrace{\boxed{(1 + D_{\chi^2}(d^{\pi_*} \| d^{\mathrm{off}})^{\frac{1}{2}}) \cdot \epsilon_{\mathrm{R,T}}}}_{\text{Offline Representation Learning}} + \underbrace{\boxed{C \cdot \left\| \frac{\partial}{\partial \theta} J_{\mathrm{BC},\phi}(\pi_\theta) \right\|_1}}_{\text{Downstream Imitation Learning}}$$

Previous theorem:

$$\mathrm{PerfDiff}(\pi_Z, \pi_*) \le (1 + D_{\chi^2}(d^{\pi_*} \| d^{\mathrm{off}})^{\frac{1}{2}}) \cdot \epsilon_{\mathrm{R,T}} + C \sqrt{\frac{1}{2} \underbrace{\mathbb{E}_{z \sim d_Z^{\pi_*}} [D_{\mathrm{KL}}(\pi_{*,Z}(z) \| \pi_Z(z))]}_{= \, \mathrm{const}(\pi_*, \phi) + \boxed{J_{\mathrm{BC},\phi}(\pi_Z)}}},$$

Only need to minimize gradient of the objective, not objective itself.