Generative AI for planning & control

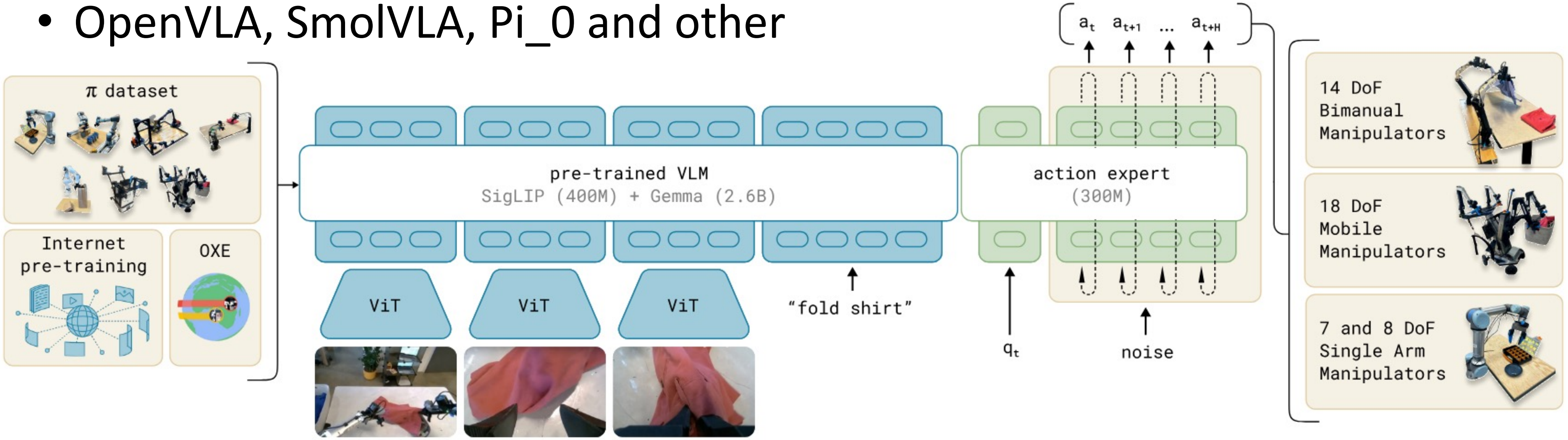
Nicolas Thome

Prof. at SU, ISIR Paris

On leave at ILLS (Mila/ETS/McGill), Montréal

Vision Language Action (VLA): the generalist

- Leverage the **high-level semantics knowledge** of LLM/VLM
- Trained with Behavioral Cloning (BC) on expert trajectories
  - OpenVLA, SmolVLA, Pi_0 and other

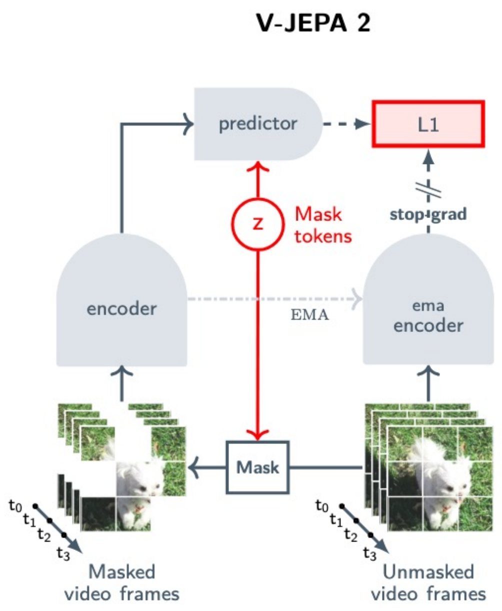


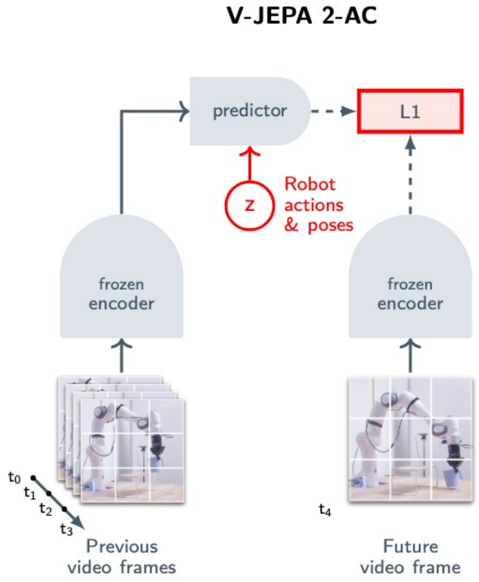
- Sota performances on several robotics / manipulation tasks
- Black box, perf. Plateau, bad in fine-grained tasks, OOD generalization

Mustafa Shukor et. al. SmolVLA: A Vision – Language – Action Model for Affordable and Efficient Robotics. ICLR 2026. Kevin Black et. al. π0: A Vision – Language – Action Flow Model for General Robot Control. RSS 2026. Moo Jin Kim et. al. OpenVLA: An Open – Source Vision – Language – Action Model. CoRL 2024.

World Models

- Predicting the evolution of the world in the input / latent space
  - V-JEPA (Meta), Genie (Google), COSMOS (NVIDIA), VaVim/VaVAM (Valeo)
- Useful for planning, especially representing rare events / corner cases
  - Relatively short horizon and simple cases

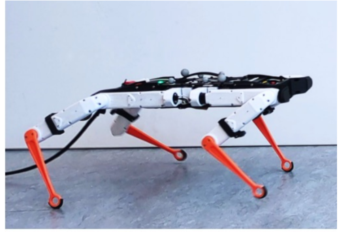




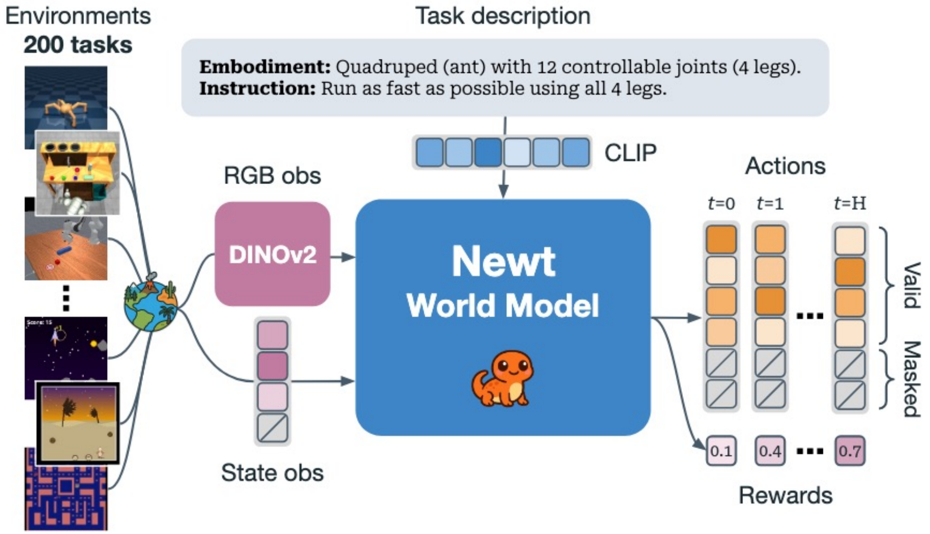
Jake Bruce et. al. Genie: Generative Interactive Environments. Arxiv, 2024.
Mido Assran et. al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. Arxiv, 2025.
Niket Agarwal et. al. Cosmos World Foundation Model Platform for Physical AI. Arxiv, 2025.
Florent Bartoccioni et. al. VaViM and VaVAM: Autonomous Driving through Video Generative Modeling. Arxiv, 2025.

Reinforcement Learning (RL): the specialist

- Learning policies with RL: very effective for a specific task
  - But also computationally demanding
- Less success of VLA in using Reinforcement Learning (RL) – needs a sufficient input signal to find reward
  - Standard strategy: fine-tune LLMs/VLMs/VLAs with RL on downstream task



- RL + world model: TD-MPC (1/2), NewT
  - Large scale, text conditioned pre-training, zero-shot generalization
  - But limited results in image-based planning

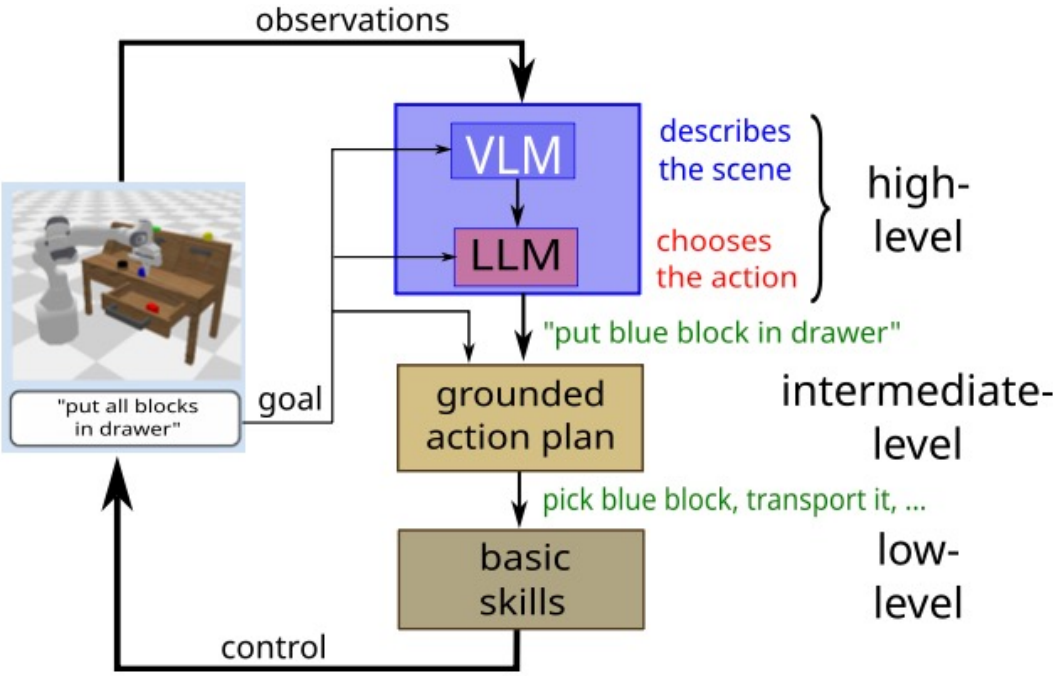


Maciej Stepien et. al. Latent Conditioned Loco-Manipulation Using Motion Priors. IEEE-RAS 24th International Conference on Humanoid Robots, 2025
Nicklas Hansen et. al. Learning Massively Multitask World Models for Continuous Control. ICLR 2026

Planning and control with AI: modular approach

- High-level planning with LMMs/VLMs: what is the best strategy for fine-tuning with RL?
  - Prompt overfitting and solutions in textual environments
  - Perception/reasoning interleaved in visual environments

- Low-level control:
  - Learning physics-informed world models (WM)
  - Diffusion Policy (DP)-MPC



Outline

1. GenAI for Planning
2. Learning physics-informed WM

Textual instruction based planning

<u>Goal:</u> clean some tomato and put it on countertop.

Textual Description
You open the fridge 1. The fridge 1 is open. In it, you see a egg 1, a mug 1, a tomato 1,.....

Planner

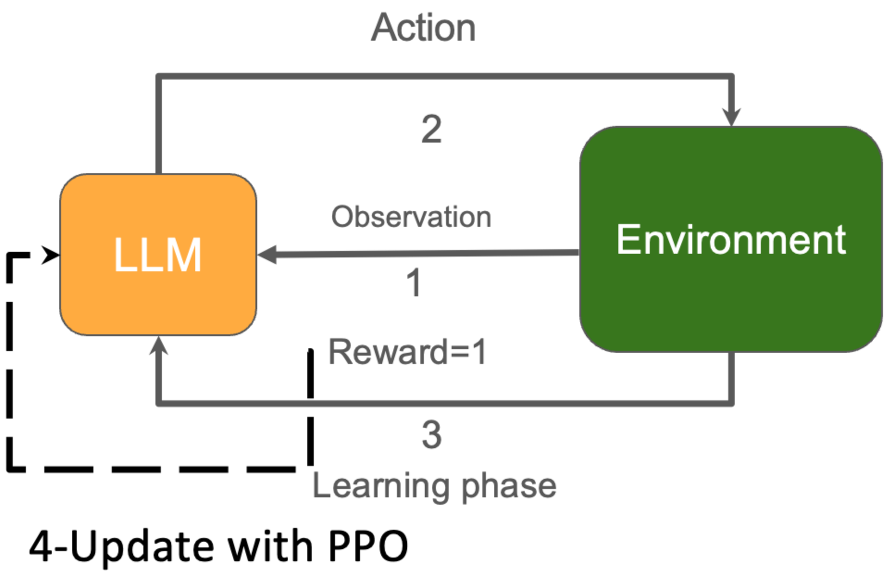
Plan:

- Go to fridge 1
- Open fridge 1
- Take tomato 1 from fridge 1
- Close fridge 1
- Go to sinkbasin 1
- Clean tomato 1 with sinkbasin 1
- Go to countertop 1
- Put tomato 1 in/on countertop 1

Textual instruction based planning

- **Leveraging Large Language Models (LLMs),** internal knowledge
- **<u>BUT:</u> grounding issues in embodied agents:**
  - Common sense
  - Learning objects physical properties, affordances
  - Interactions between objects, causality

- Recent works on fine-tuning embodied agents with RL, e.g., PPO
- **Open questions:**
  - Generalization
  - Catastrophic forgetting



T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.Y. Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. ICML 2023
W. Tan, W. Zhang, S. Liu, L. Zheng, X. Wang, and B. An. True knowl-dge comes from practice: Aligning llms with embodied environments via reinforcement learning. arXiv, 2024.

Our study: prompt overfitting protocol

Prompt Strategy

P_0:
Possible actions of the agent: close fridge, close kitchen cupboard, close oven, take bottle of cold water from kitchen cupboard, take clean mug from dining table
Goal: clean the Kitchen
Observation: You can see a fridge. Empty! You can see an opened kitchen cupboard. The kitchen cupboard contains a bottle of cold water. Oh, great. Here's an oven. The oven is empty, You lean against the wall, inadvertently pressing a secret button. The wall opens up to reveal a dining table. On the dining table you see a clean mug.
Inventory: You are carrying nothing.
Next action of the agent:

P_1:
Goal: clean the Kitchen
Inventory: You are carrying nothing.
Observation: You can see a fridge. Empty! You can see an opened kitchen cupboard. The kitchen cupboard contains a bottle of cold water. Oh, great. Here's an oven. The oven is empty, You lean against the wall, inadvertently pressing a secret button. The wall opens up to reveal a dining table. On the dining table you see a clean mug.
Possible actions of the agent:'close fridge', 'close kitchen cupboard', 'close oven', 'take bottle of cold water from kitchen cupboard', 'take clean mug from dining table'
Next action of the agent:

P_2:
<Begin Possible actions> close fridge, close kitchen cupboard, close oven, take bottle of cold water from kitchen cupboard, take clean mug from dining table **<End Possible actions>**
<Begin Goal>clean the Kitchen **<End Goal>**
<Begin Observation> You can see a fridge. Empty! You can see an opened kitchen cupboard. The kitchen cupboard contains a bottle of cold water. Oh, great. Here's an oven. The oven is empty, You lean against the wall, inadvertently pressing a secret button. The wall opens up to reveal a dining table. On the dining table you see a clean mug.**<End Observation>**
<Begin Inventory> You are carrying nothing. **<End Inventory>**
Next action :

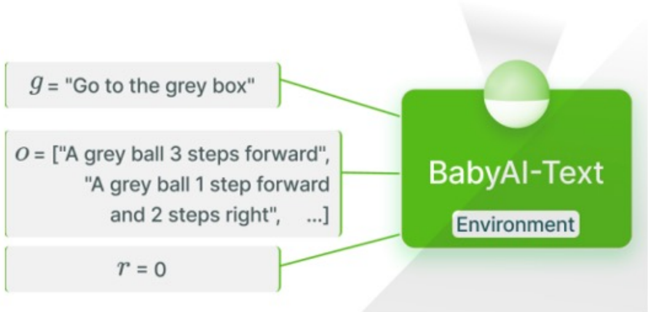
P_3 :
Welcome to TextWorld! You find yourself in a messy house. Many things are not in their usual location. Let's clean up this place. After you'll have done, this little house is going to be spick and span! Look for anything that is out of place and put it away in its proper location. What you can do is to close fridge, close kitchen cupboard, close oven, take bottle of cold water from kitchen cupboard, take clean mug from dining table. Your goal is to clean the Kitchen. You can see a fridge. Empty! You can see an opened kitchen cupboard. The kitchen cupboard contains a bottle of cold water. Oh, great. Here's an oven. The oven is empty, You lean against the wall, inadvertently pressing a secret button. The wall opens up to reveal a dining table. On the dining table you see a clean mug.. Now, You are carrying nothing., and your next action is to

Switch in order

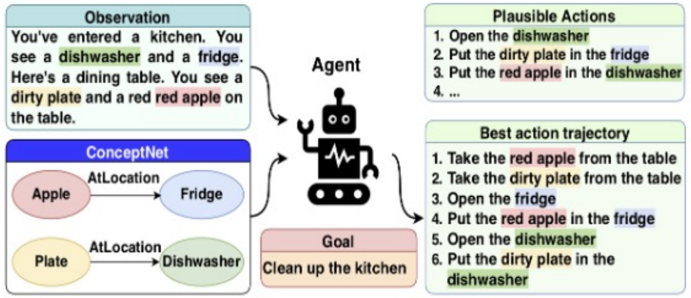
Rigid syntaxe

Paraphrase Natural Language

Environments

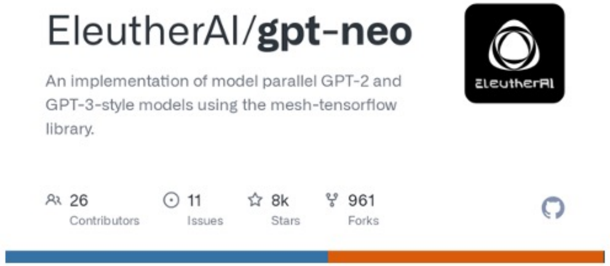


Baby Ai Text: requires exploration and understanding objects' positions



Text World Common sense: requires commonsense knowledge about the world.

Training & Evaluation Scenarios



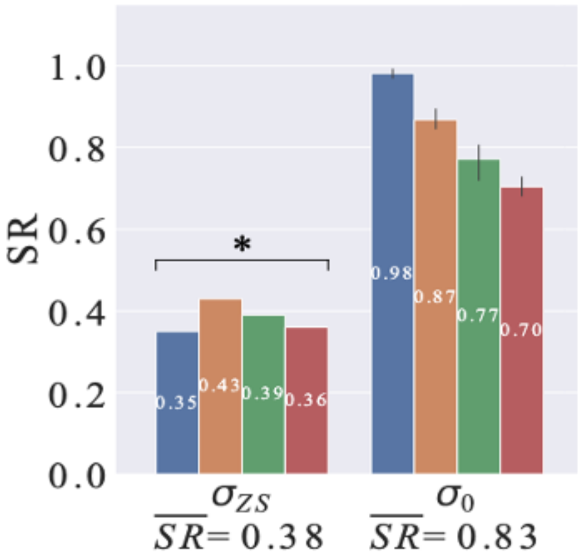
- Zero shot evaluation
- Train with one Strategy and Evaluate with others
- Train on All Strategy

Experimental results: prompt sensitivity

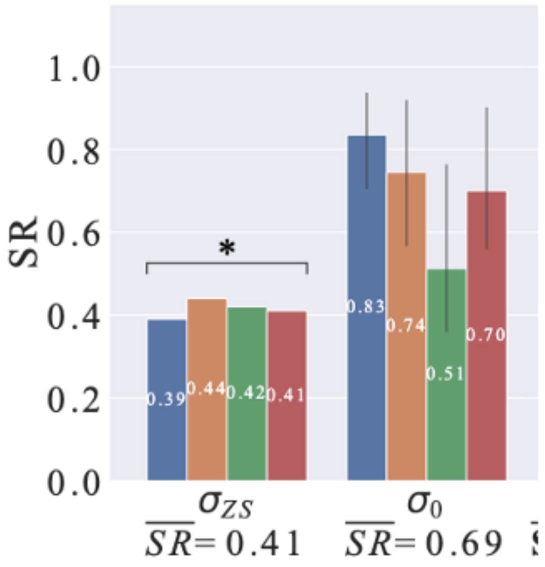
Metrics:

- Success Rate (SR): $SR = \frac{n_e}{N}$
- Mean Episode Length

Results for FLAN T5 780M



TWC Medium



BabyAI Text

- RL boost performances
- **But strong overfitting to the prompt**
- Similar trend with GPT-neo

- Mitigating prompt overfitting with contrastive learning

Visual instruction based planning

Goal: clean some tomato and put it on countertop.

Text **riptio**
You op e
fridge 1 is you see
a egg o
1,.....

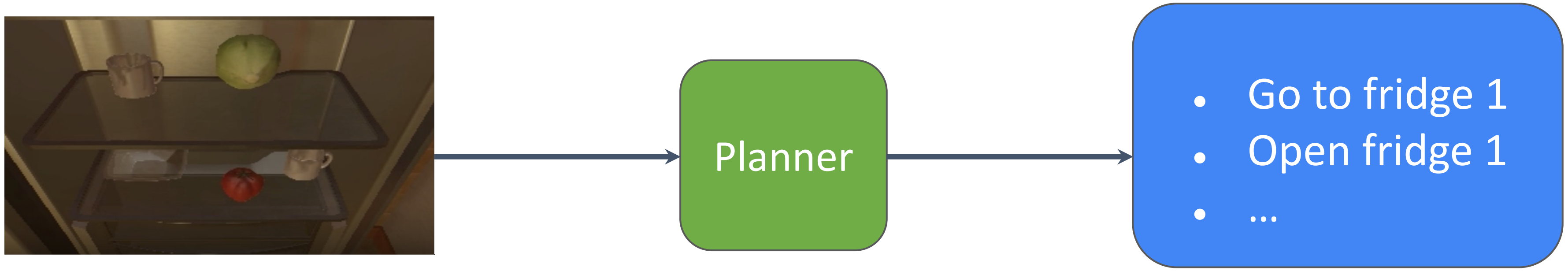
Planner

Plan:

- Go to fridge 1
- Open fridge 1
- Take tomato 1 from fridge 1
- Close fridge 1
- Go to sinkbasin 1
- Clean tomato 1 with sinkbasin 1
- Go to countertop 1
- Put tomato 1 in/on countertop 1

Visual instruction based planning: AlfWorld

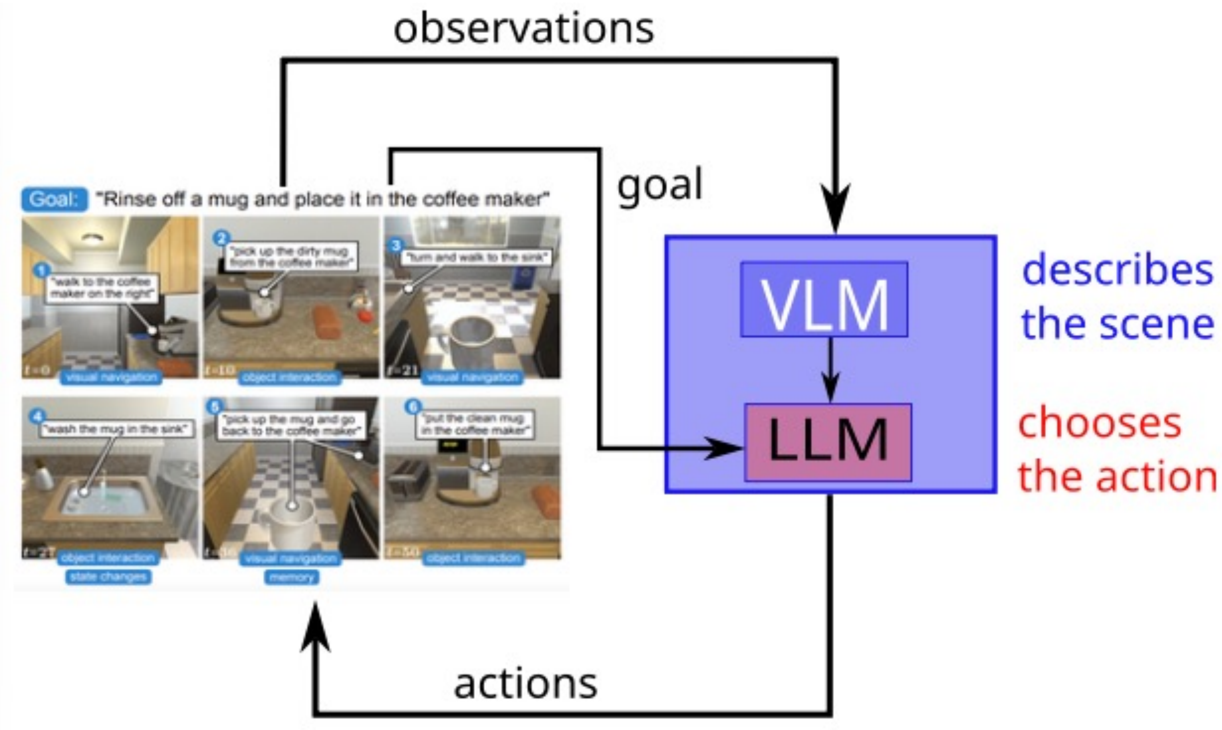
- Both image and detailed textual descirption of each image
- Sucessufull attempts for using LLMs, either with RL or filtering relevant actions
- However, performances with VLMs from visual inputs are way less sucessfull, e.g., limited performances in recent works RL4VLM
    - EMMA: textual guidance, cumbersome and unrealistic requirement

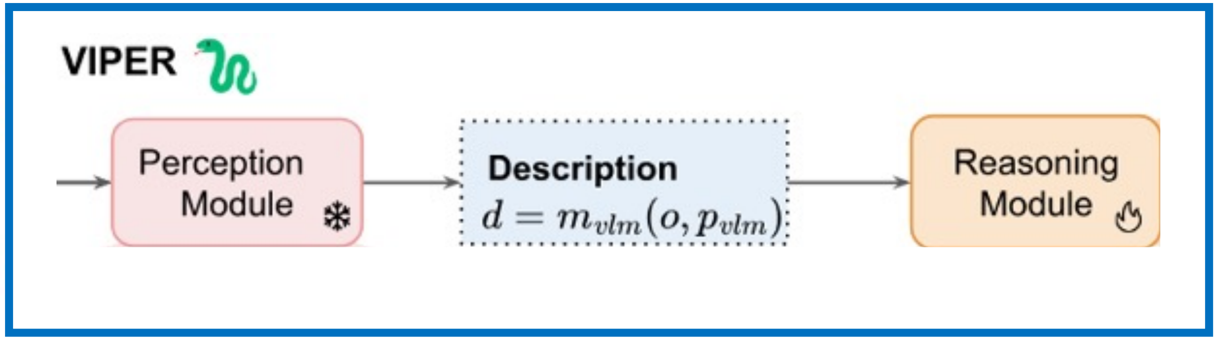


Y. Zhai, H. Bai, Z. Lin, J. Pan, S. Tong, Y. Zhou, A. Suhr, S. Xie, Y.LeCun, Y. Ma, S. Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. NeurIPS 2024.
Y. Yang, T. Zhou, K. Li, D. Tao, L. Li, L. Shen, X. He, J. Jiang, Y. Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. CVPR 2024.

VIPER: Visual Perception and Explainable Reasoning for Sequential Decision-Making

Separate perception (VLM) and decision making (LLM)

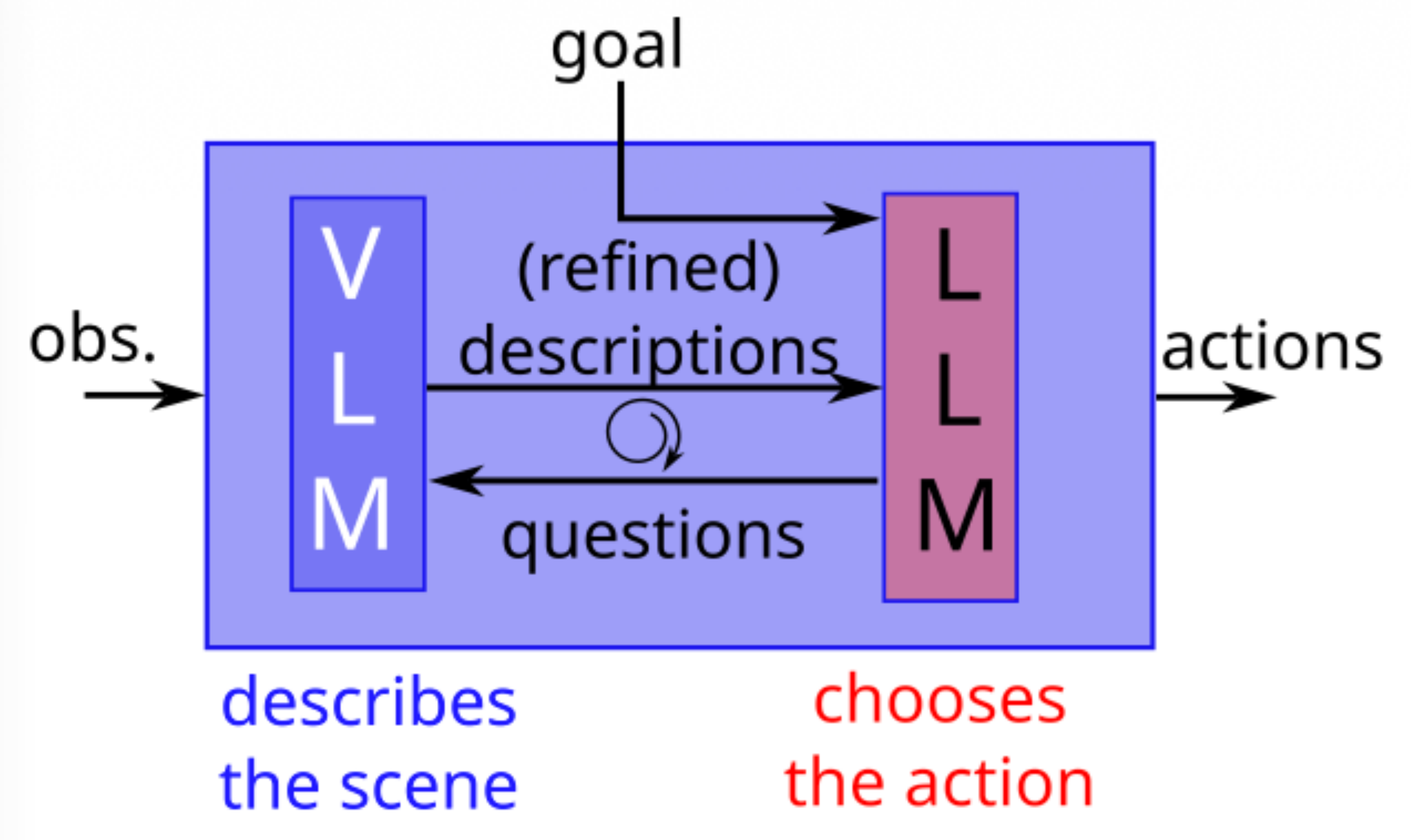




- VLM: zero-shot (ZS), LLM fine tuned with BC+RL
- Intermediate text => explainability / monitoring potential

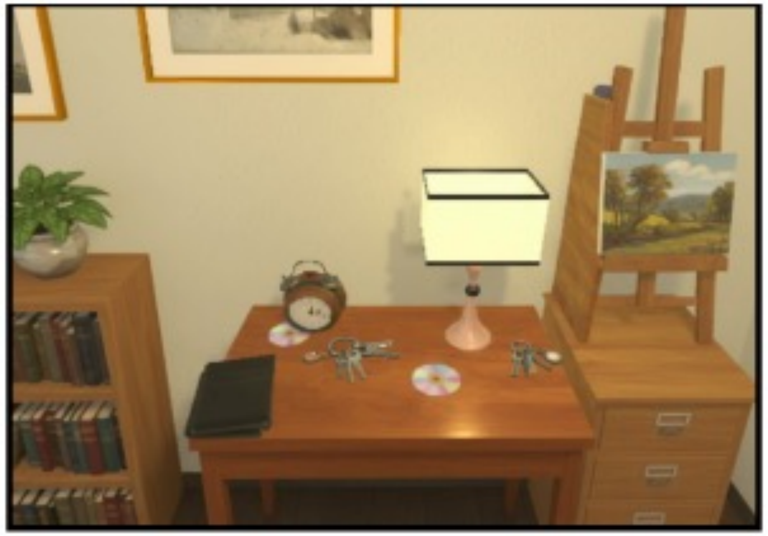
Extension: Perception and Reasoning Interleaved for Sequential Decision Making (PRISM)

Key insight (cognitive sciences): perception should be goal- and action-oriented



<u>Idea:</u> keep ZS VLM, but add a question-answering mechanism between LLM/VLM to iteratively refine perception

PRISM example



Goal: find a keychain and put it in drawer.

w/o Interleaved Perception Reasoning

Description: In this picture we can see a table, there is a clock, lamp and some other things present on the table, in the background we can find a wall.

Action: go to countertop 1

with Interleaved Perception Reasoning

Do you see any keychains on the table? Yes, there are two keys attached to small discs sitting near the lamp base.
Do you see a drawer nearby? Yes, there is a drawer nearby. It is located on the right side of the image

Description: In this scene, a table holds a clock, a lamp, and two keychains attached to small discs, which are situated near the lamp base. There is also a drawer on the right side of the image.

Action: Take keychain 1 from sidetable

PRISM results on ALFWORLD

Supervision	Method	Agent type	Pick↑	Look↑	Clean↑	Heat↑	Cool↑	Pick2↑	Avg ↑
Access to text env	ReAct	LLM	71%	28%	65%	62%	44%	35%	54%
	DEPS	LLM	93%	100%	50%	80%	**100%**	00%	76%
	Reflexion	LLM	**96%**	**94%**	**100%**	81%	83%	88%	**91%**
	EMMA	VLM	71%	88%	94%	85%	83%	67%	82%
	EMAC+	VLM	79%	88%	93%	**90%**	90%	74%	86%
No access to text env	MiniGPT-4	VLM	04%	17%	00%	19%	17%	06%	16%
	RL4VLM	VLM	47%	14%	10%	14%	18%	18%	21%
	Idefics2 8B*	VLM	75%	77%	74%	69%	67%	50%	69%
	VIPER$_{BC}$	VLM+LLM	**80%**	77%	67%	87%	**71%**	53%	72%
	VIPER$_{BC+RL}$	VLM+LLM	**80%**	77%	77%	**92%**	**71%**	53 %	75%
	PRISM$_{BC}$	VLM+LLM	**80%**	77%	81%	87%	**71%**	62%	77%
	PRISM$_{BC+RL}$	VLM+LLM	**80%**	**83%**	**84%**	**92%**	**71%**	**67%**	**80%**

- Access to text env: upper bound
- Separating LLM/VLM vs training a VLM: ~ performances
- QA adds a large gain, RL furthers boost performances

Separation LLM/VLM: model monitoring

- Detecting important tokens : Integrated Gradient (IG)
- Action to description: identify part of the description important for action
- For important token, highlight important regions, detect connected component in image => **weakly-supervised instance segmentation**
- Important text token <-> important image regions: powerful tool to analyse success and failure modes (perception vs reasoning)

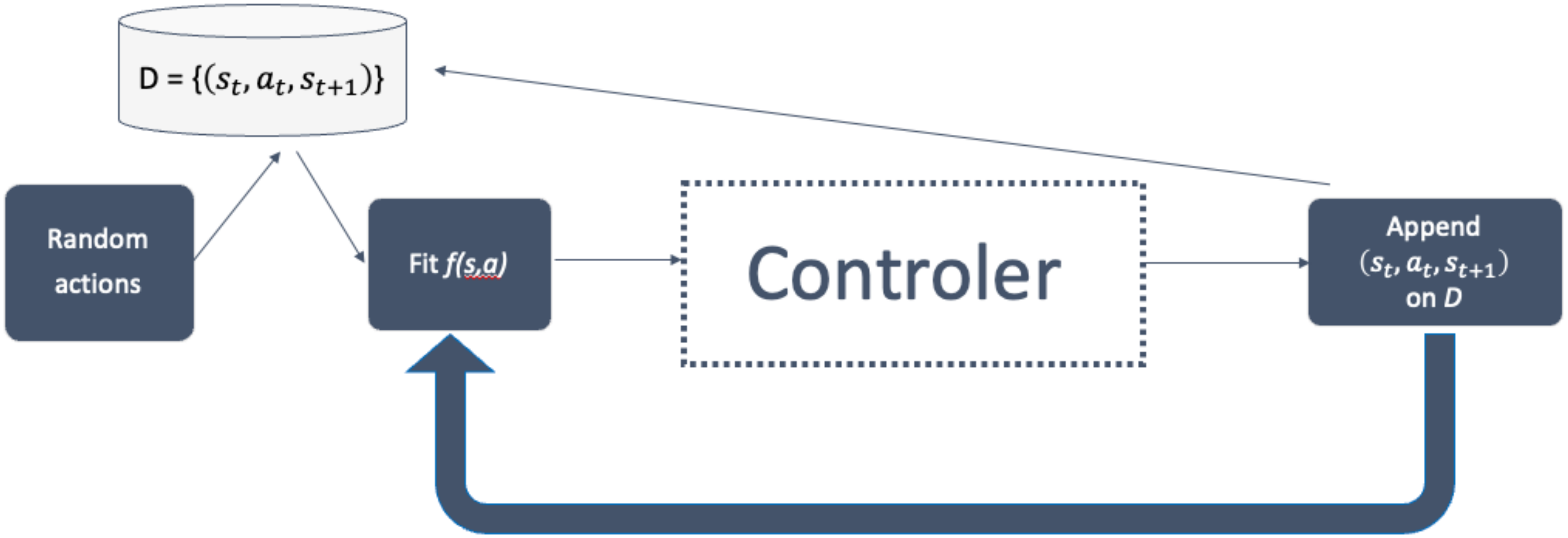
Outline

1. GenAI for Planning
2. Learning physics-informed WM

Model based RL, e.g., PETS (2018) [1]

- Learning an world (dynamical) model $f(s_t,a_t) = s_{t+1}$ - or $p(s_{t+1} | s_t,a_t)$
  - Supervised learning
  - Using random transition
  - Or follwing some policy

- Control through $f(s_t,a_t)$ to choose actions



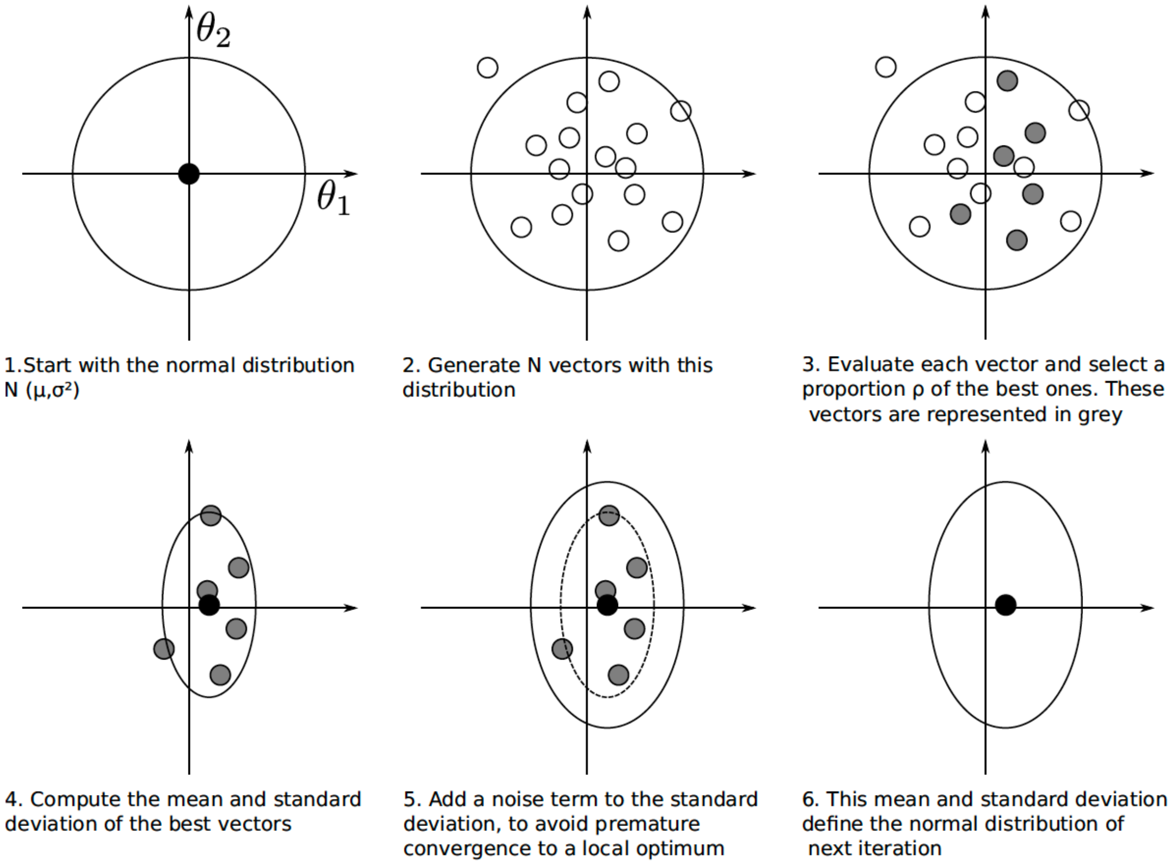
[1] K. Chua, R. Calandra, R. McAllister, and S. Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. NeurIPS 2018.

Model Predictive Control (MPC) [2]

- From existing $f(s_t,a_t)$, Control to choose actions

- Linear $f(s_t,a_t)$ and quadractic reward
  : linear–quadratic regulator (LQR)

- More general: cross-entropy methods (CEM) [3]
  - Iterate 1->6 until convergence

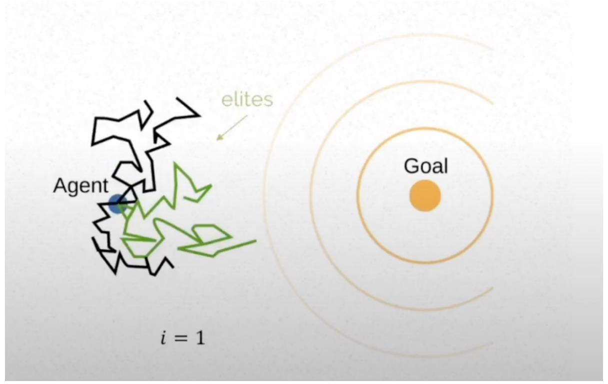


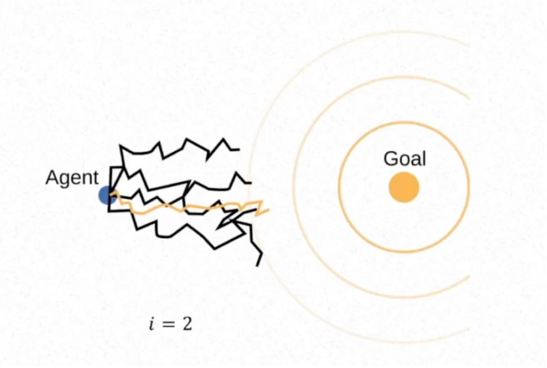
[2] J. Richalet, A. Rault, J. Testud, J. Papon. Model Predictive Heuristic Control: Applications to Industrial Processes. *Automatica*, 1978.
[3] R.Y. Rubinstein, D.P. Kroese. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning, Springer-Verlag, New York, 2004.

CEM for sequential decision making [4]

- Horizon H, cumulative reward

$$\boldsymbol{A}_t^{(H)} = \underset{\boldsymbol{A}_t^{(H)}}{\operatorname{argmax}} \sum_{t'=t}^{t+H-1} r(\boldsymbol{s}_{t'}, \boldsymbol{a}_{t'})$$





- **Main CEM hyper-parameters**
  - Number of iterations T
  - Sample size N
  - Horizon H

PETS

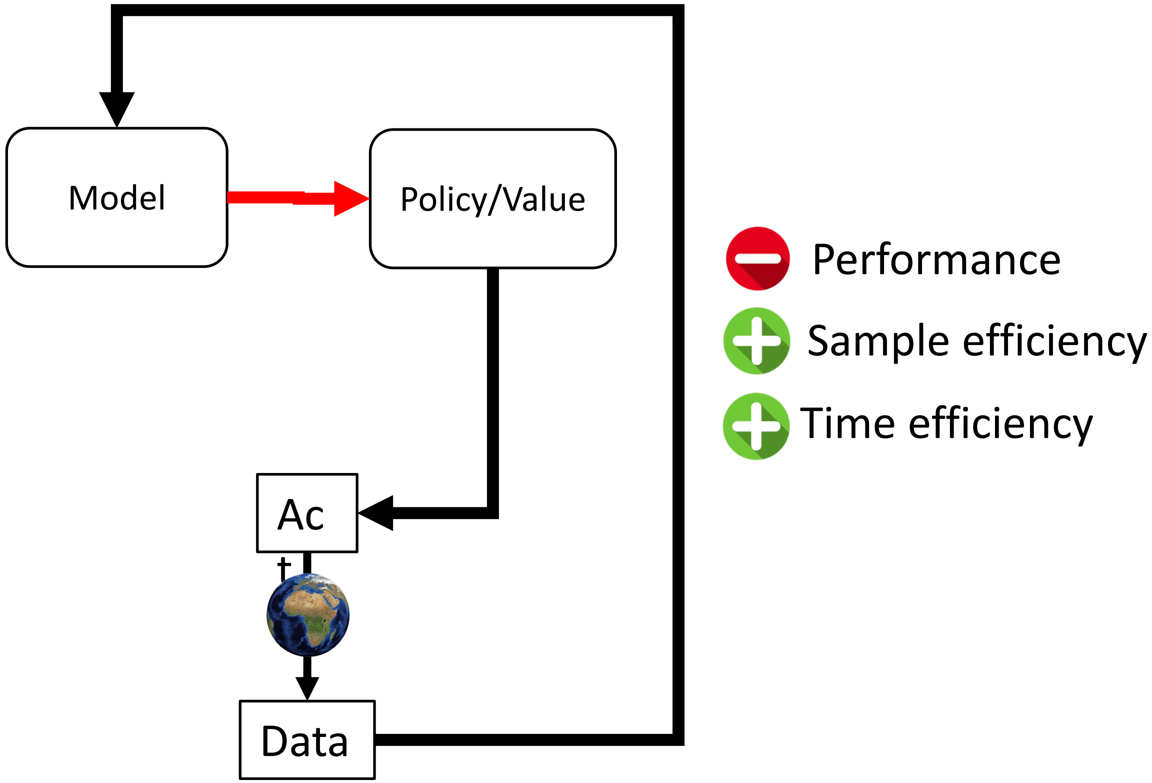
- (+) Sample efficiency: learning f(s,a) OK
- (+) Asymptotic performance
- (−) Time efficiency

Large T, N, H!

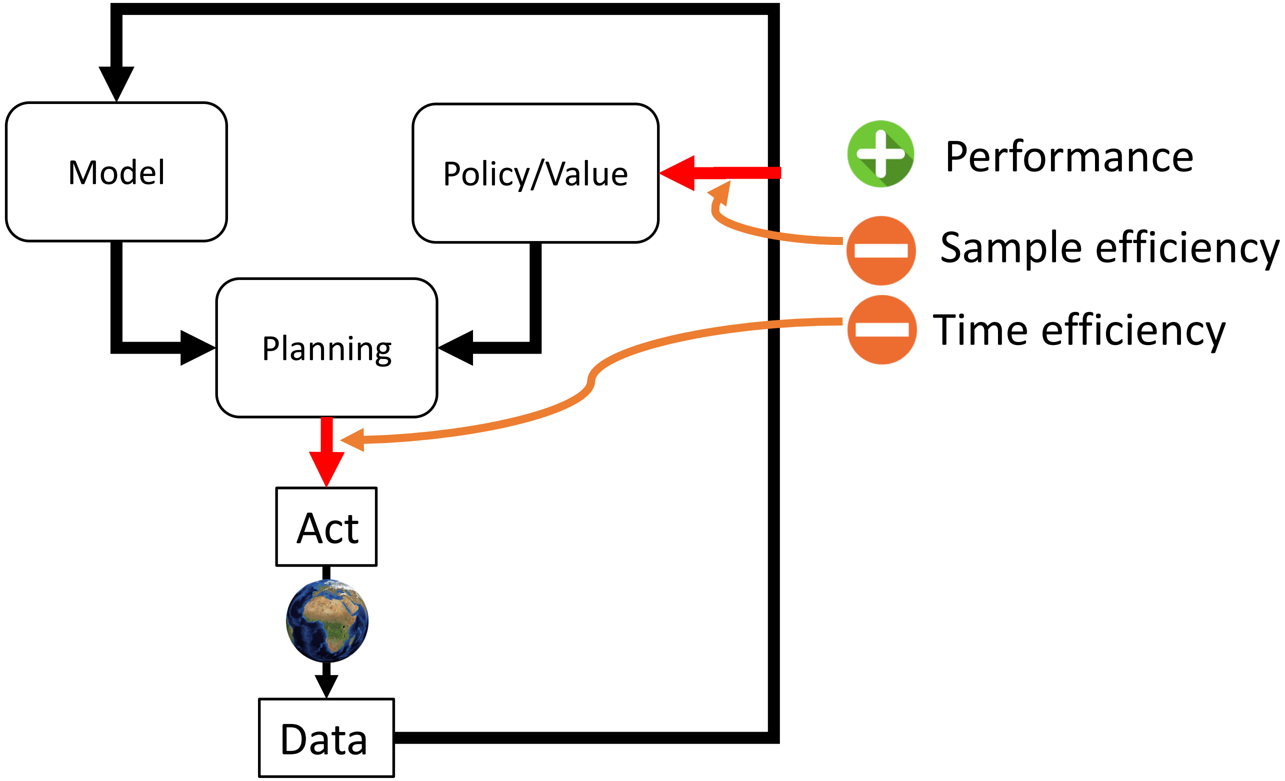
[4] C. Pinneri, S. Sawant, S. Blaes, J. Achterhold, J. Stueckler, M., Rolinek, G.Martius. Sample-efficient Cross-Entropy Method for Real-time Planning. Conference on Robot Learning, 2020.

CEM improvements

- Dyna style RL [5]: learn f(s,a), then learn a policy in imagination



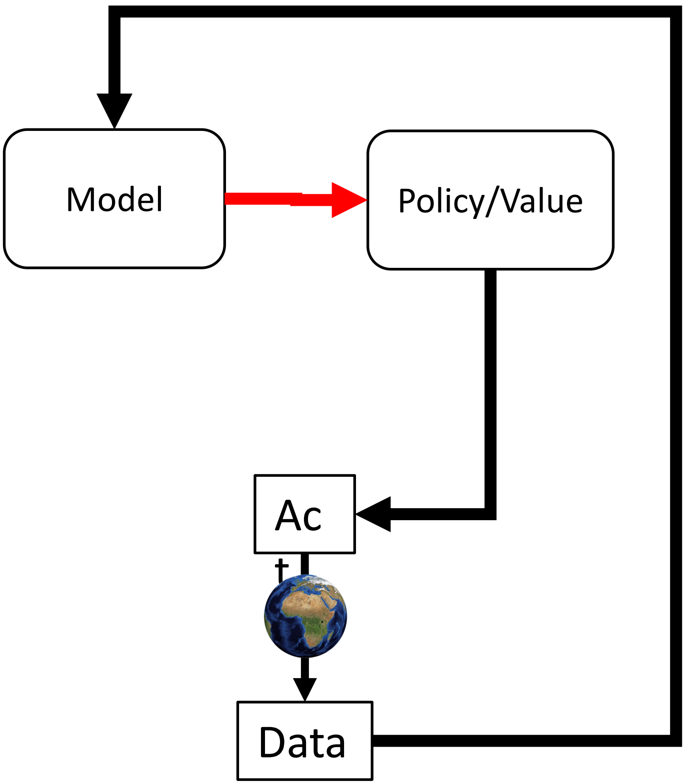
- TD-MPC [6]: learn model-free policy to speed-up MPC



[5] Harshit Sikchi, Wenxuan Zhou, and David Held. Learning off-policy with online planning. CoRL 2022.
[6] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. ICML, 2022.

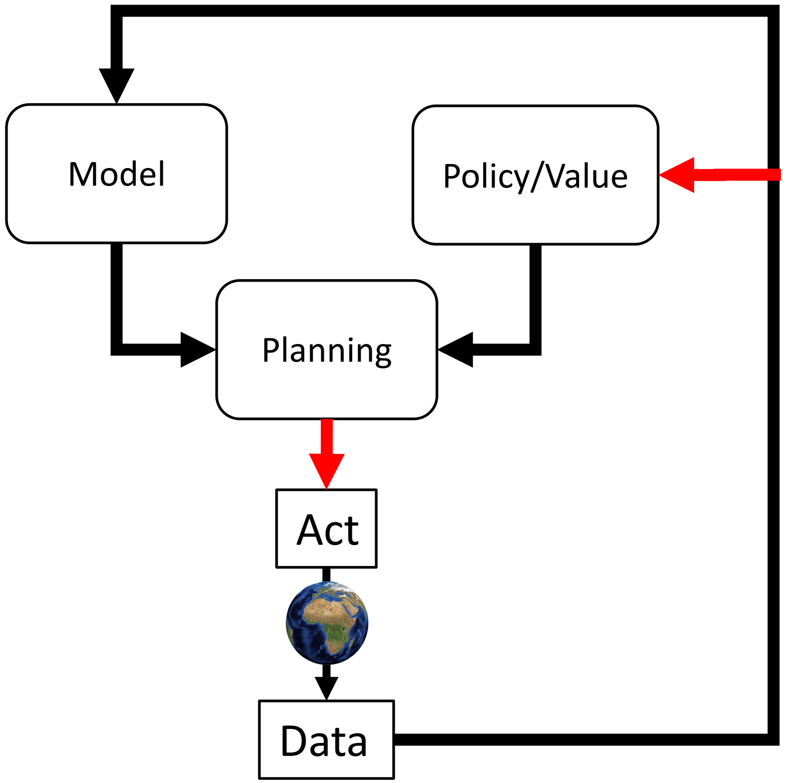
Physics-Informed Model and Hybrid Planning (PhIHP)

Dyna style RL [5]



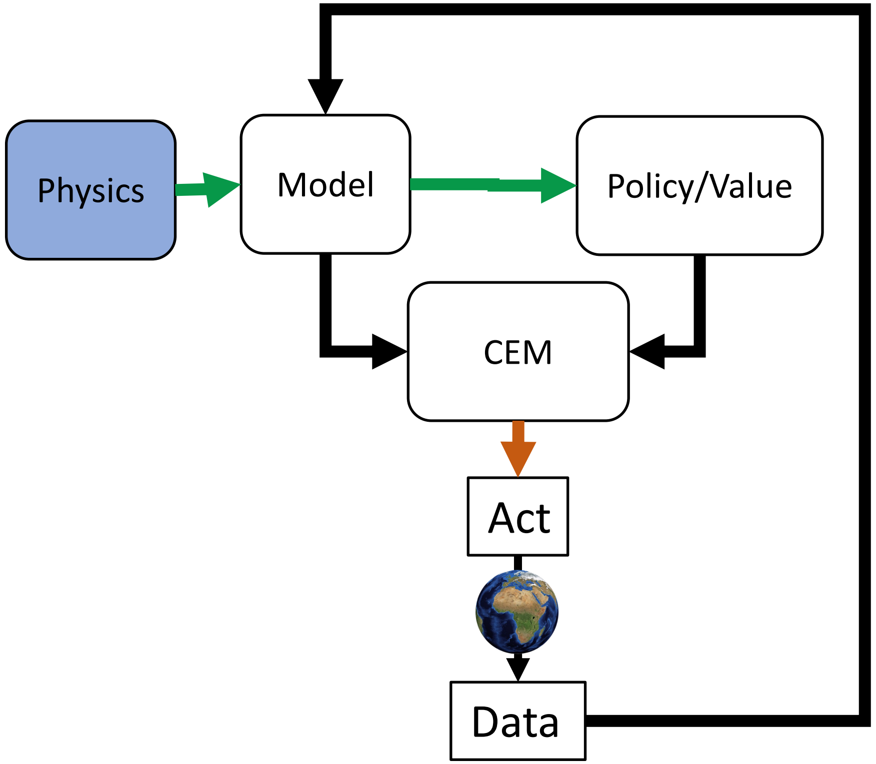
- Asymptotic performance
- Sample efficiency
- Time efficiency

TD-MPC [6]



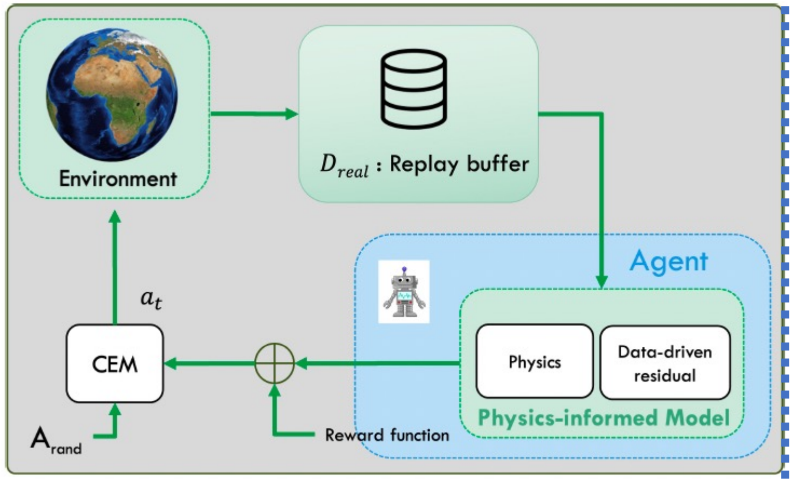
- Asymptotic performance
- Sample efficiency
- Time efficiency

PhIHP (Ours)



- Asymptotic performance
- Sample efficiency
- Time efficiency

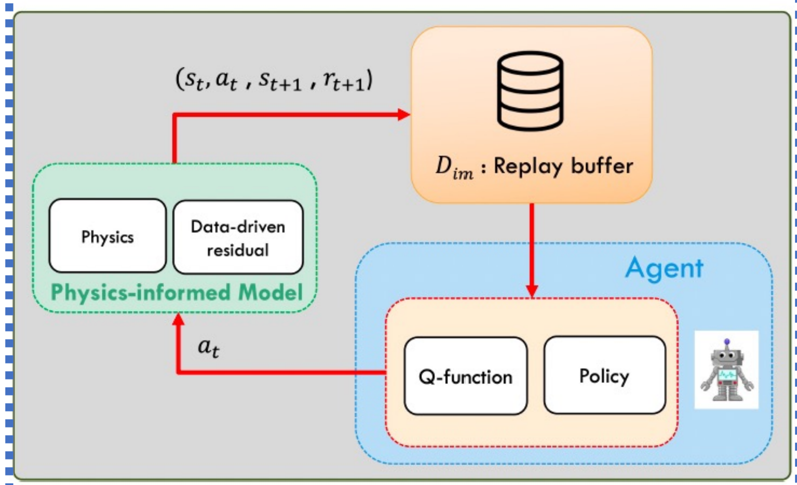
PhIHP pipeline



(a) Learn a physics-informed model

- Sample efficiency
- Performance (↓ bias)

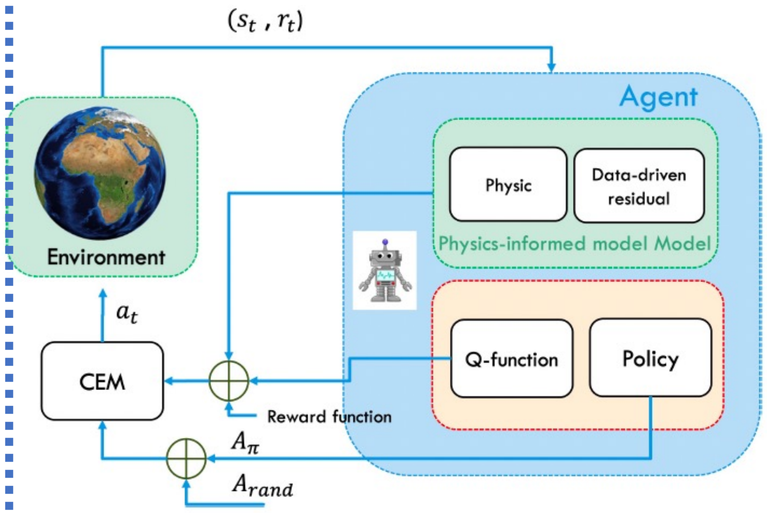
Hybrid model



(b) Learn an actor/critic offline

- Sample efficiency

Learning in imagination



(c) Behaviour at inference time

Hybrid TD3/MPC controller

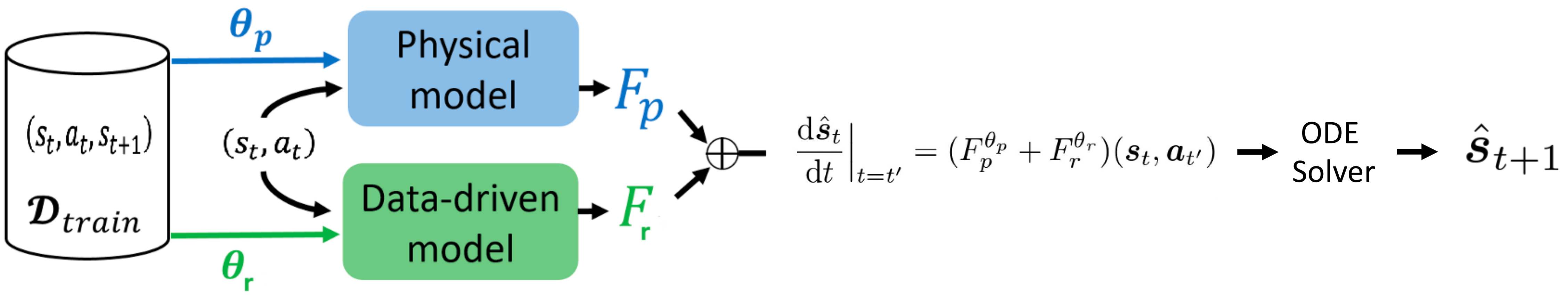
- Sample efficiency
- Time efficiency
- Asymptotic performance

PhIHP: learning a residual model

Physics: an approximate model described as ODE [7]
Learning residual + physical parameters

- Sample efficiency
- Asymptotic performance

Training Strategy:



[7] Yuan Yin, Vincent Le Guen, Jeremie Dona, Emmanuel de Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. "Augmenting physical models with deep networks for complex dynamics forecasting." ICLR, 2021.

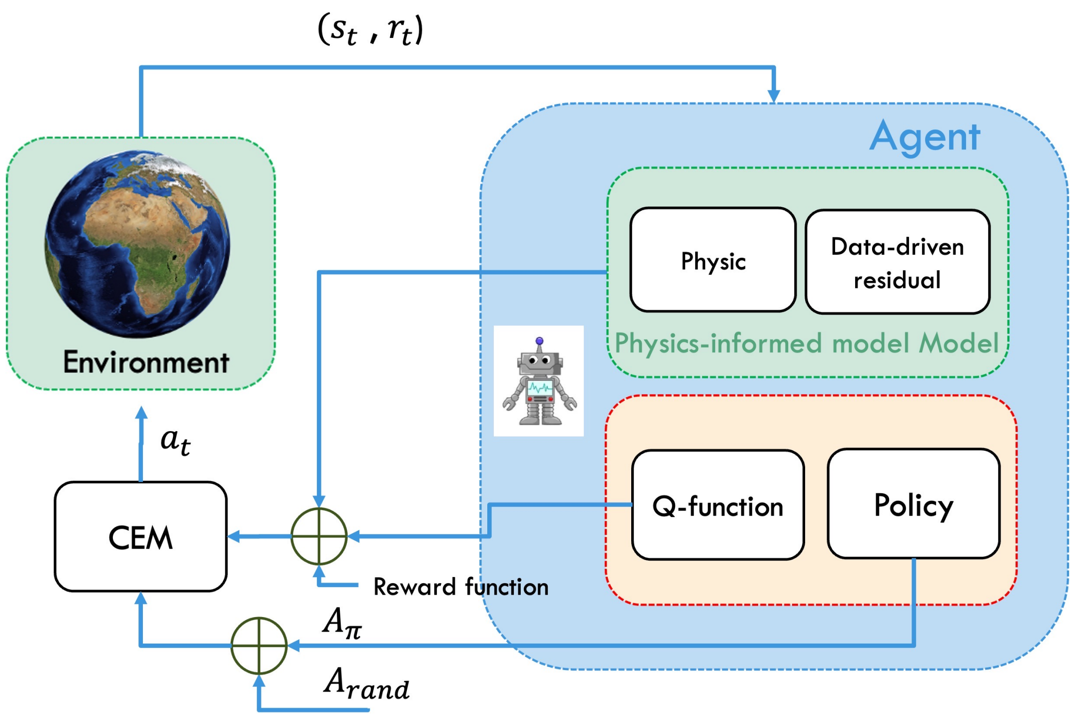
PhIHP: hybrid controller

- Combine A_{rand} (MPC) and A_{π} (TD3)
- Trajectory score

$$A^* = \arg\max_{A \in \mathcal{A}^H} \left(\sum_{t=t_0}^{H} \gamma^{t-t_0} R(s_t, a_t) + \alpha \cdot \gamma^{H-t_0} Q(s_H) \right)$$

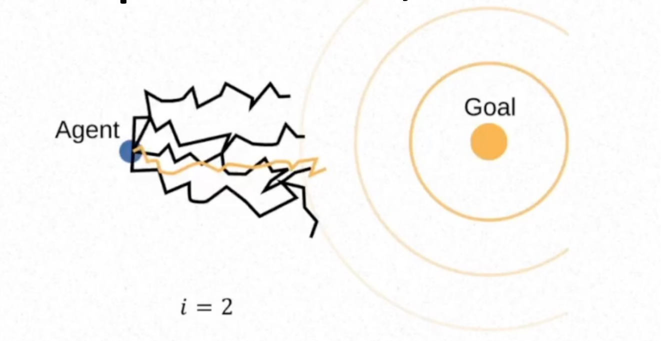
local solution

long-term reward



- **Good policy π + good physics-informed model**

=> optimizing horizon H, sample size N, # iterations T



- Time efficiency
- Asymptotic performance

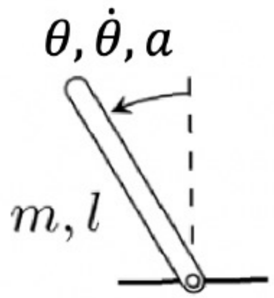
Experiments & results

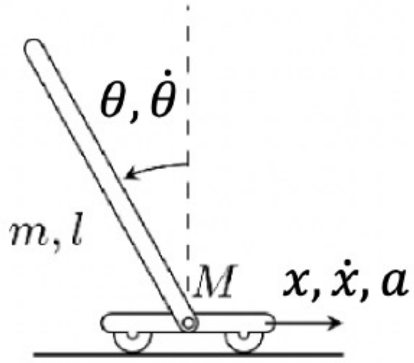
Approximate physics: no friction

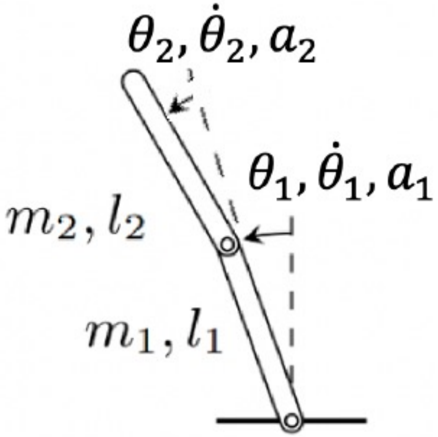
Pendulum

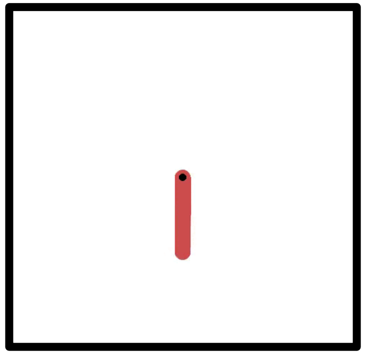
CartPole

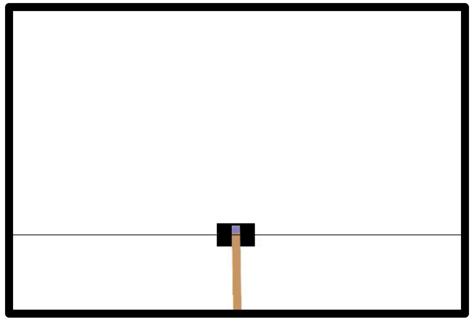
Acrobot

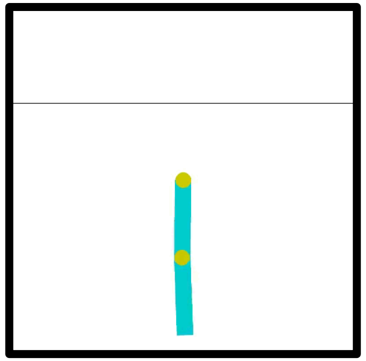










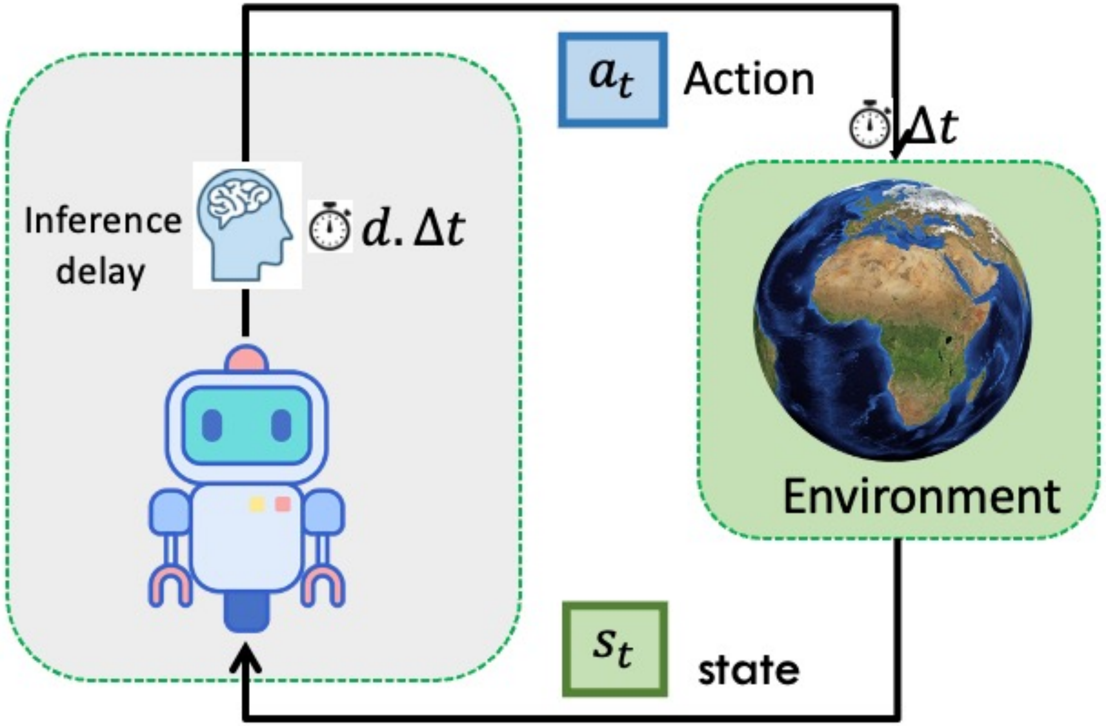


Real-Time Hybrid control with Physics (RT-HCP)

Extending PhiP to directly learn on robotic platforms

<u>**Main challenge:**</u> **inference delay in embedded devices**

- **d steps**

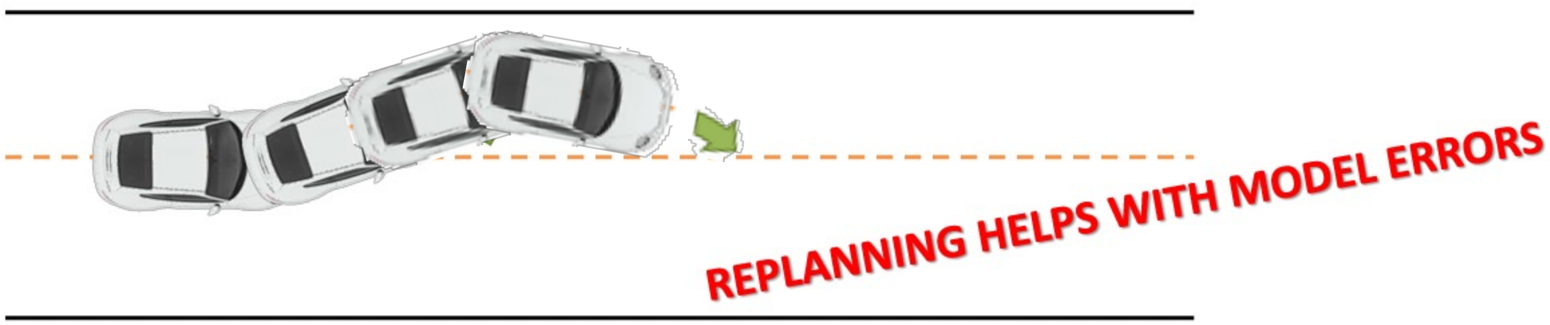


Env state	s_0	...	s_d	...	s_{2d}	...	$s_{d.t}$
Env action	a_0		a_1		a_2		a_t

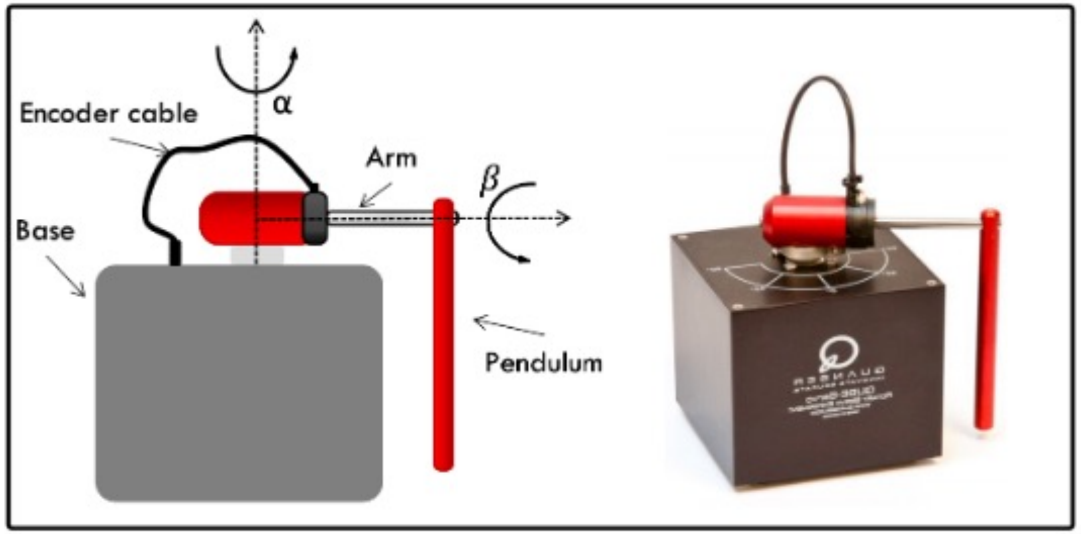
Methodology: how many actions to send?

- Δt given, H^p : Planning Horizon – setup for desired performances
- Compute inference time T^i for MPC with H^p
- **General idea:** sending the minimal number of actions before re-planning

$$H^e_{min} = int\left(\frac{T^i}{\Delta t}\right) + 1$$

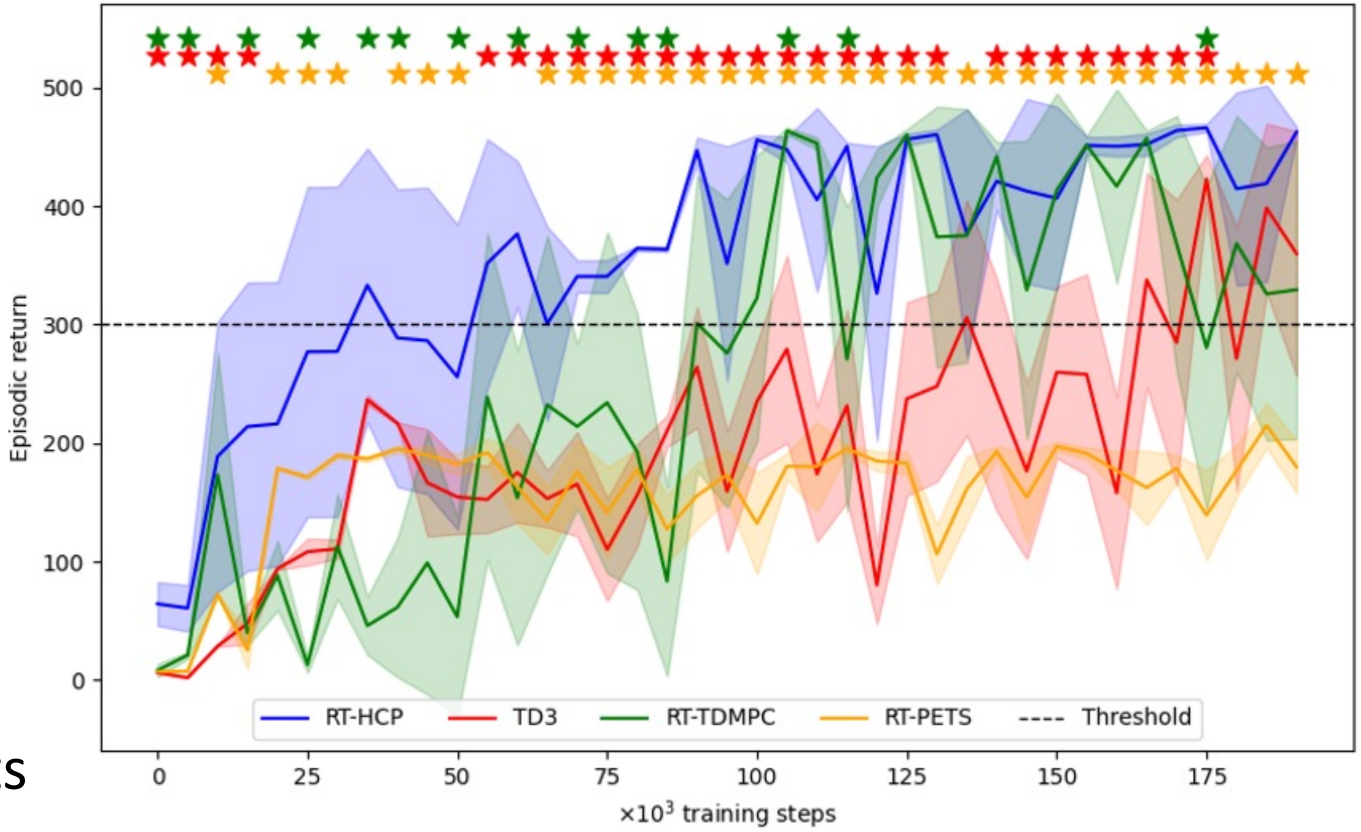


Experiments & results



Real Furuta pendulum

- Approximate model: double pendulum
  - Fine-tuning physical parameters
- Learning residual friction and cable effects

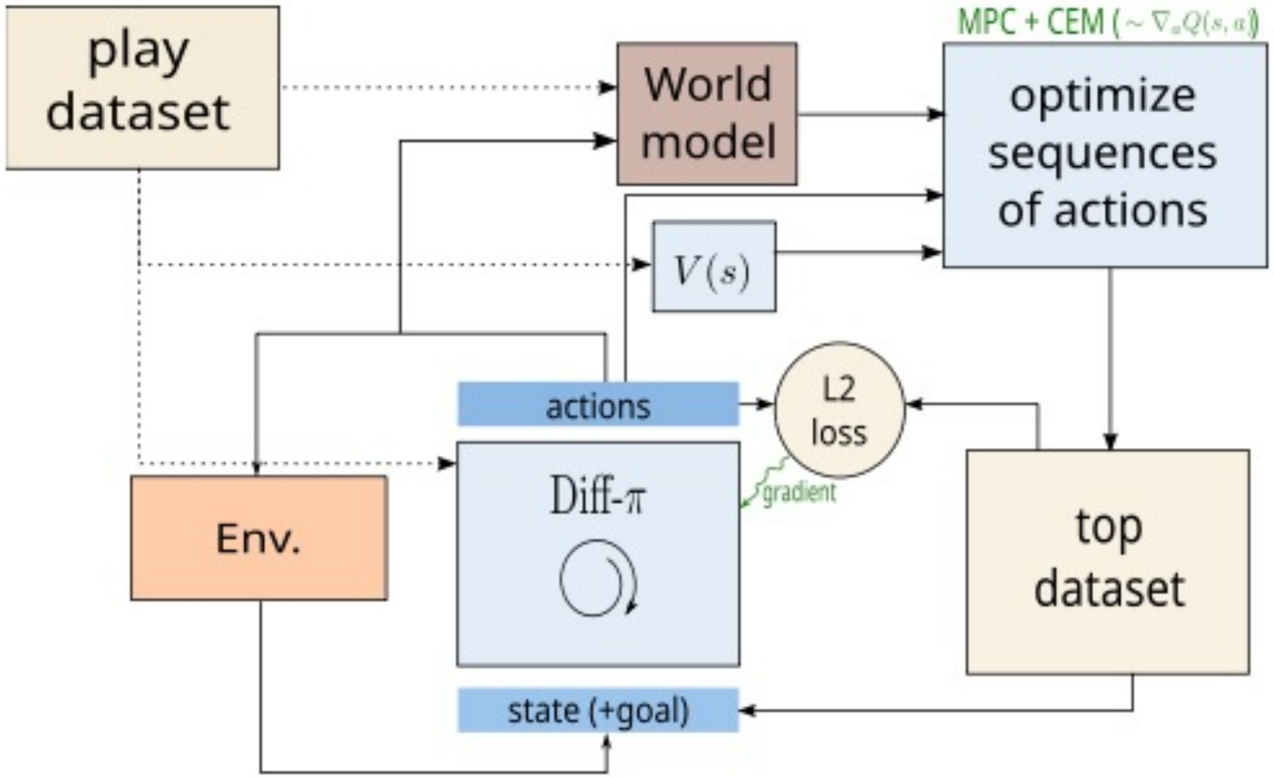


Results

	RT-HCP	RT-TDMPC	TD3	RT-PETS
Fine-tuning physical parameters 60k steps 23 minutes				
100k steps 35 minutes				
160k steps 58 minutes				

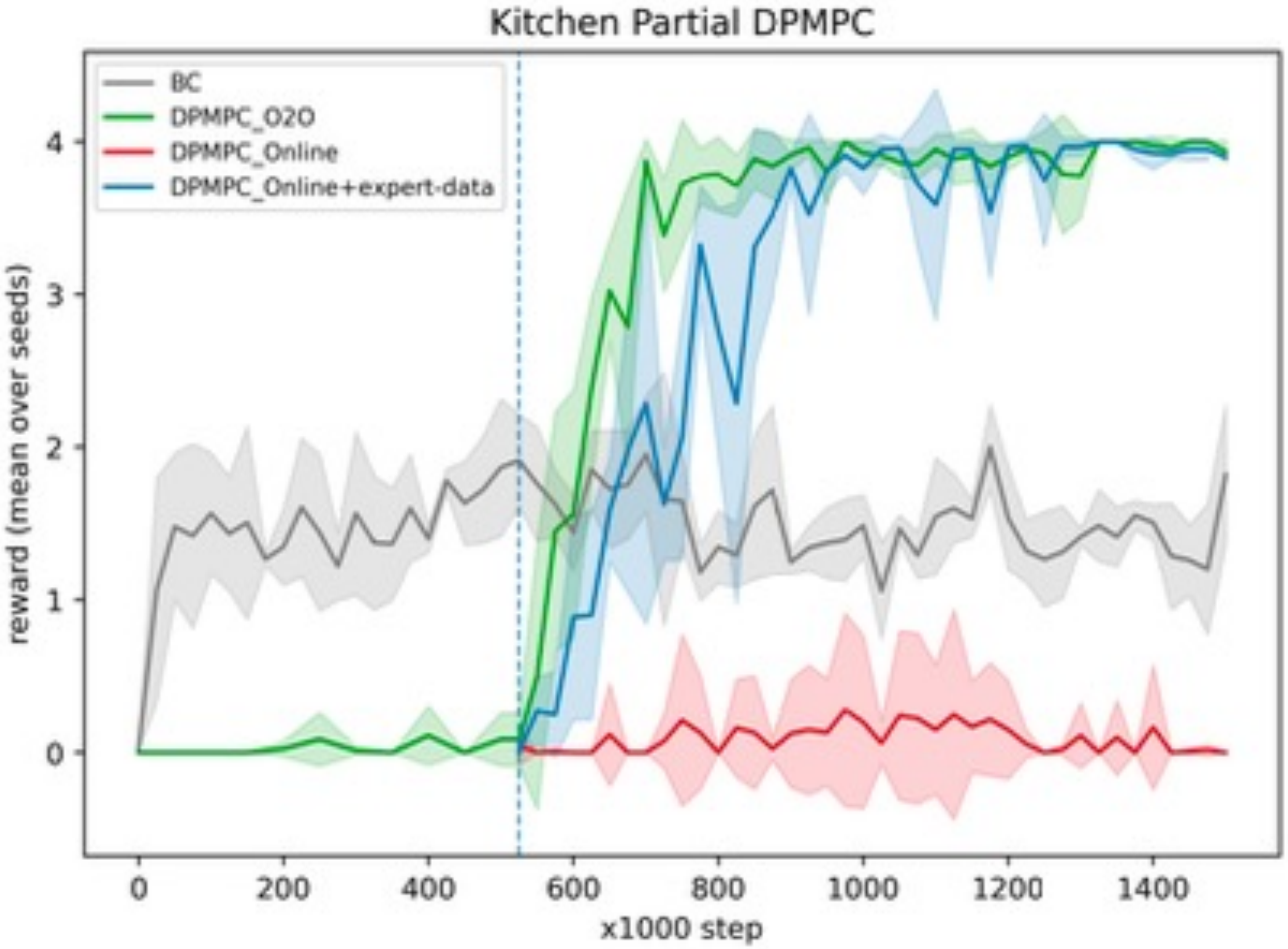
Current work: Diffusion Policy (DP)-MPC

- TD-MPC with DP vs TD3/SAC
- Diffusion policy trained MPC guidance (+ RL)
  - Main idea: distill MPC -> learned policy



Diffusion Policy (DP)-MPC: preliminary results

- DP-MPC > BC
- Using a pre-trained DP helps
- Offline to Online (O2O) vs Online + expert data



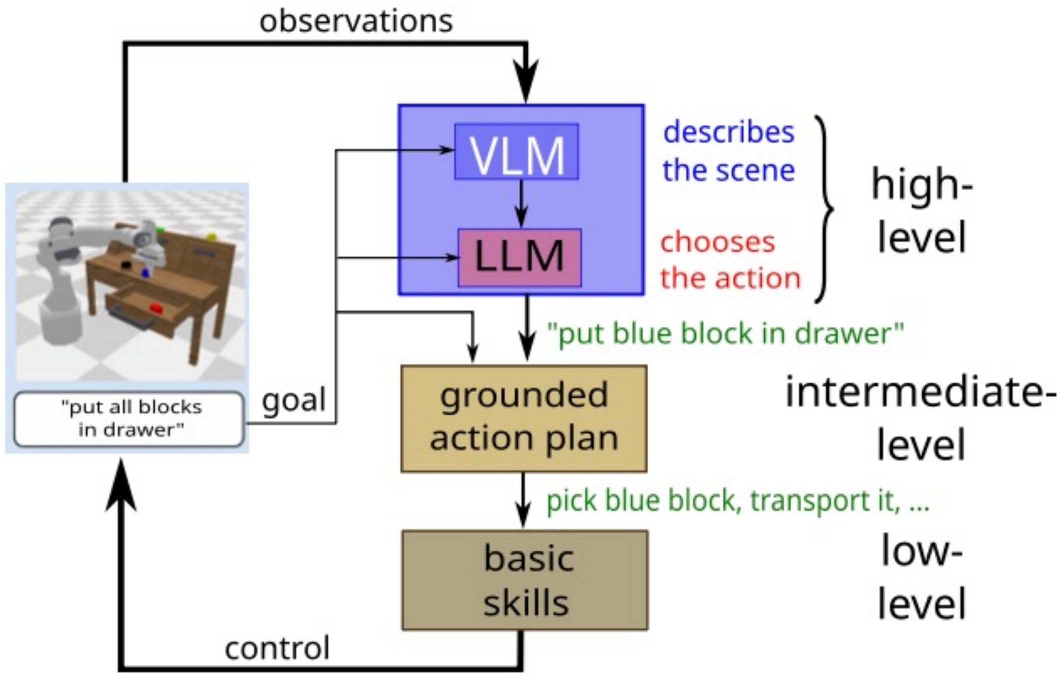
Conclusion & perspectives

Physics-informed world model

- Combining physical priors for proprioception + image encoder
- More complex approximate dynamics, needs differentiable simulator

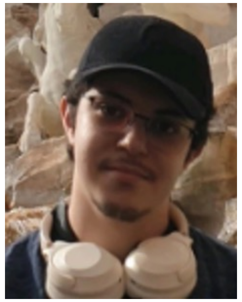
Combination of high- and low-level

- PRISM + DP
  - vs VLA + CoT, CoT / action
- WM for action verification / grounding
  - Uncertainty => re-planning



Thank you for your attention!

Zakariae El Asri



Salim Aissi

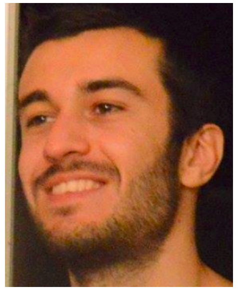
Clémence Grislain

Clément Romac

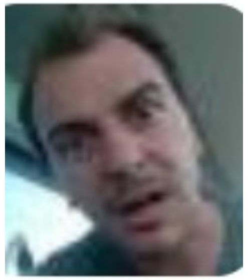
Thomas Carta

Laure Soulier

Olivier Sigaud



Clément Rambour



Sylvain Lamprier

Pierre-Yves Oudayer

[EST24] Physics-Informed Model and Hybrid Planning for Efficient Dyna-Style Reinforcement Learning. Z. El Asri, O. Sigaud, N. Thome. **RLC 2024.**

[ERS+25] RT-HCP: Dealing with Inference Delays and Sample Efficiency to Learn Directly on Robotic Platforms. Z. El Asri, O. Sigaud, N. Thome. **IROS 2025.**

[ERL+22] Residual Model-Based Reinforcement Learning for Physical Dynamics. Z. El Asri, C. Rambour, V. Le Guen, N. Thome. **NeurIPS 2022 Offline RL Workshop.**

[ARC+24] Reinforcement Learning for Aligning LLM Agents: Quantifying and Mitigating Prompt Overfitting.
S. Aissi, C. Romac, T. Carta, S. Lamprier, P.Y. Oudeyer, O. Sigaud, L. Soulier, N. Thome. **NAACL 2025 Findings.**

[AGS+26] Perception Reasoning Interleaved for Sequential Decision Making.
S. Aissi, C. Grislain, L. Soulier, M. Chetouani, O. Sigaud, N. Thome. **Under review at ICML 2026.**

[AGS+25] Reinforcement Learning for Aligning LLM Agents: Quantifying and Mitigating Prompt Overfitting.
S. Aissi, C. Grislain, L. Soulier, M. Chetouani, O. Sigaud, N. Thome. **Arxiv 2025.**

[CRW+23] Grounding large language models in interactive environments with online reinforcement learning
T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, P.Y. Oudeyer.. **ICML 2023.**