

Les modèles vision-langage-action

Comprendre les politiques et leur apprentissages

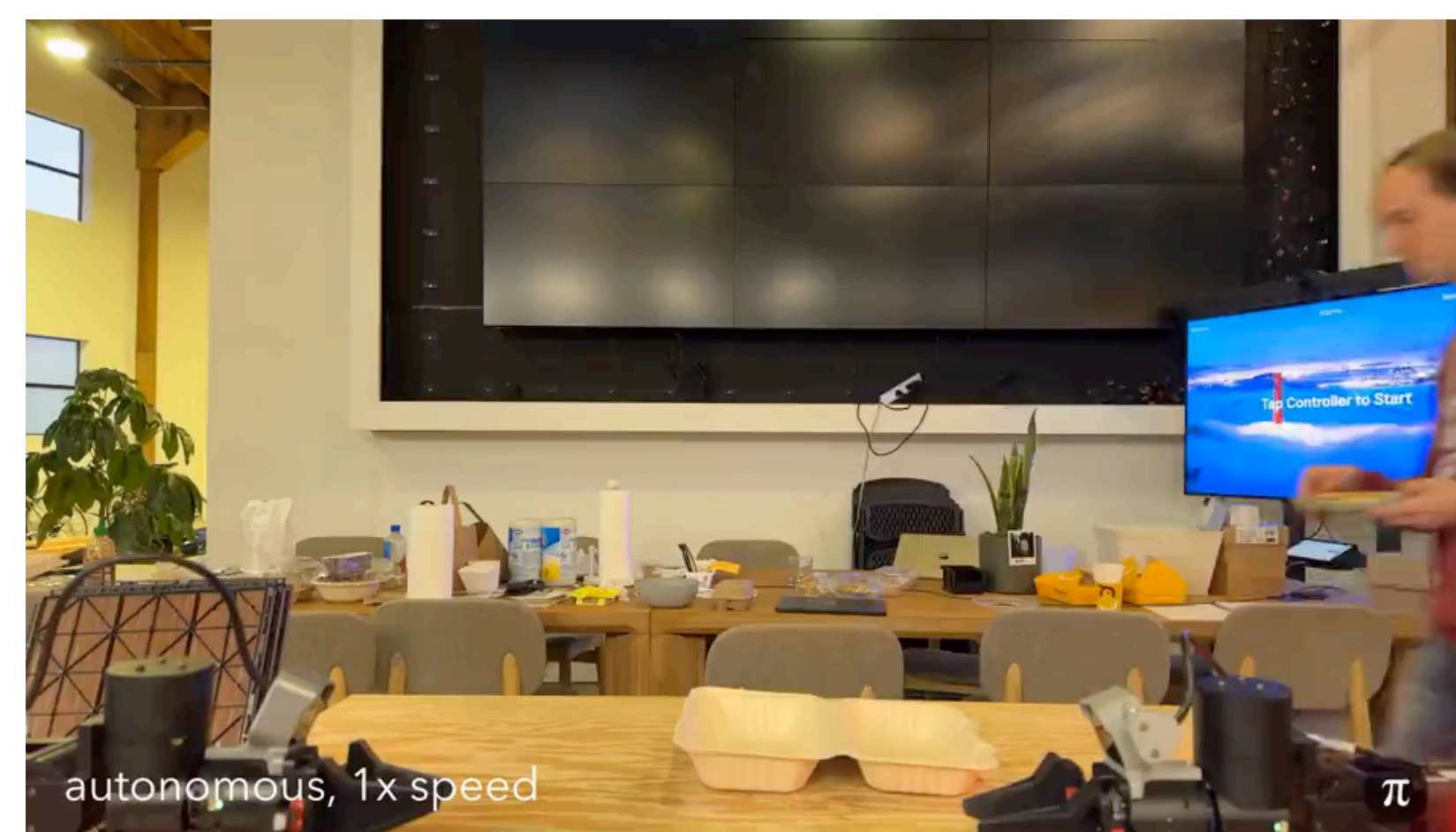
Alex Zongo

Doctorant, Systèmes Intelligent Aérospatiaux



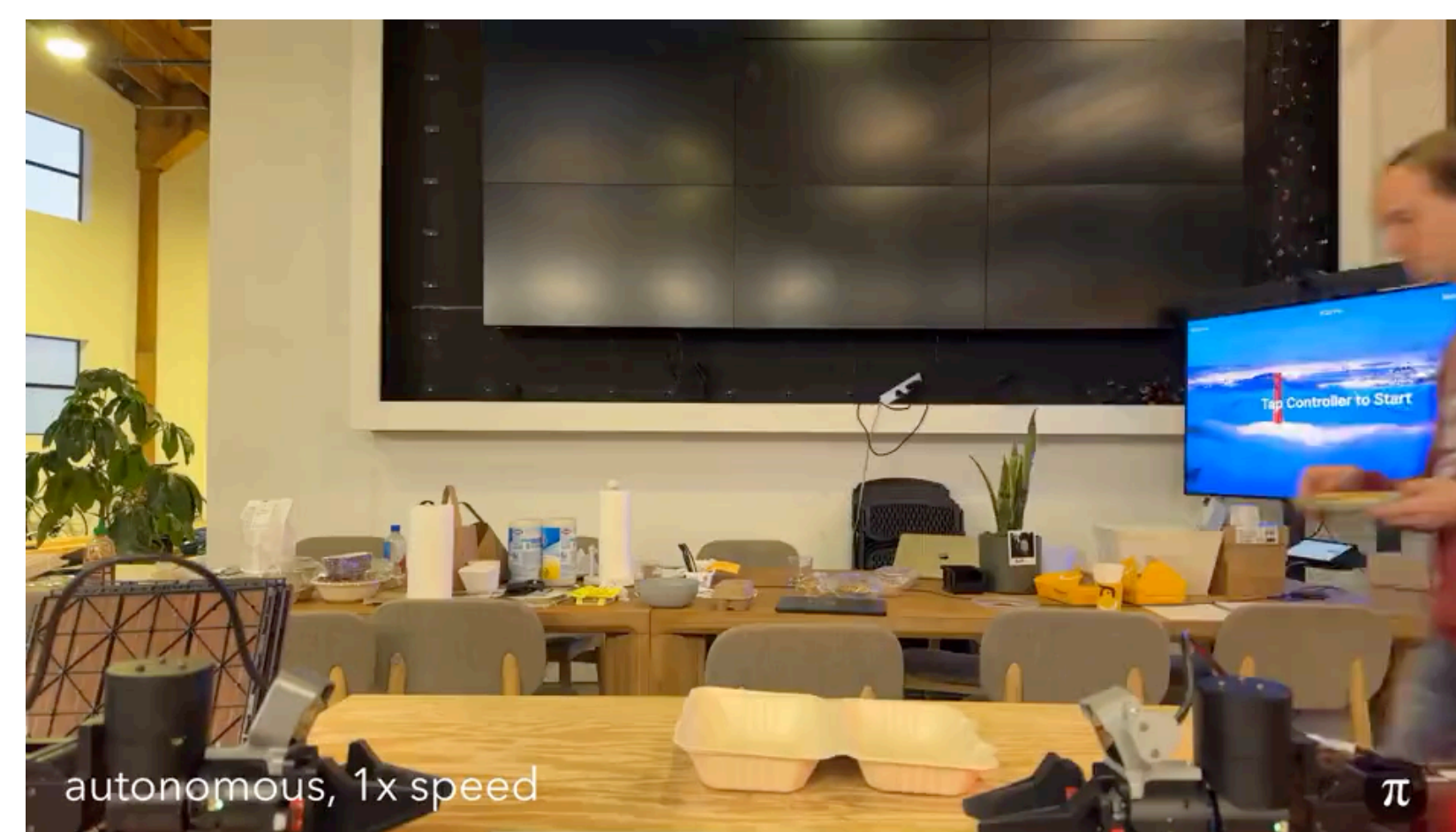
alexzongo.com





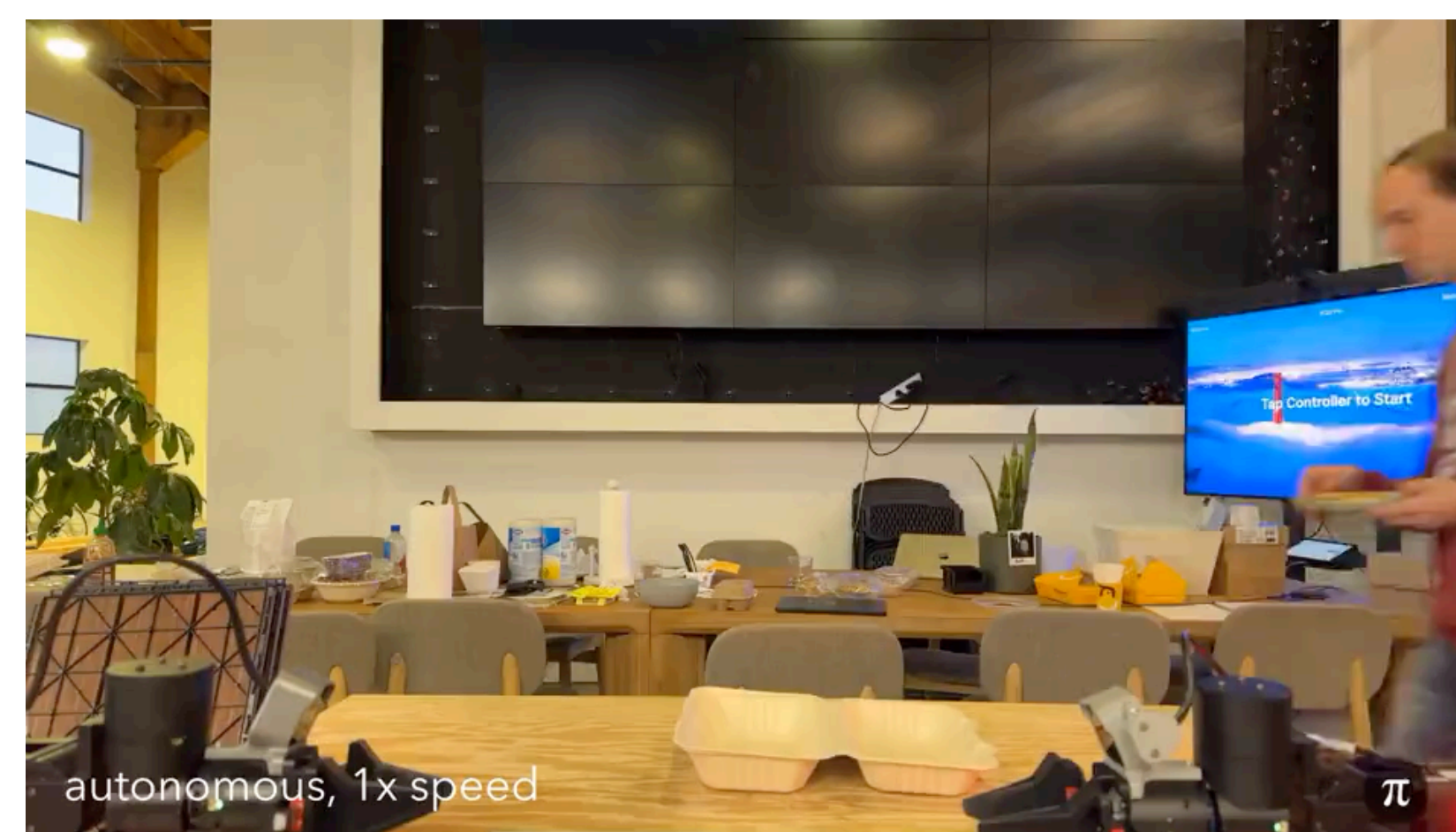
Credits: <https://www.pi.website/blog/pi0>

SEINAR · Octobre 2026



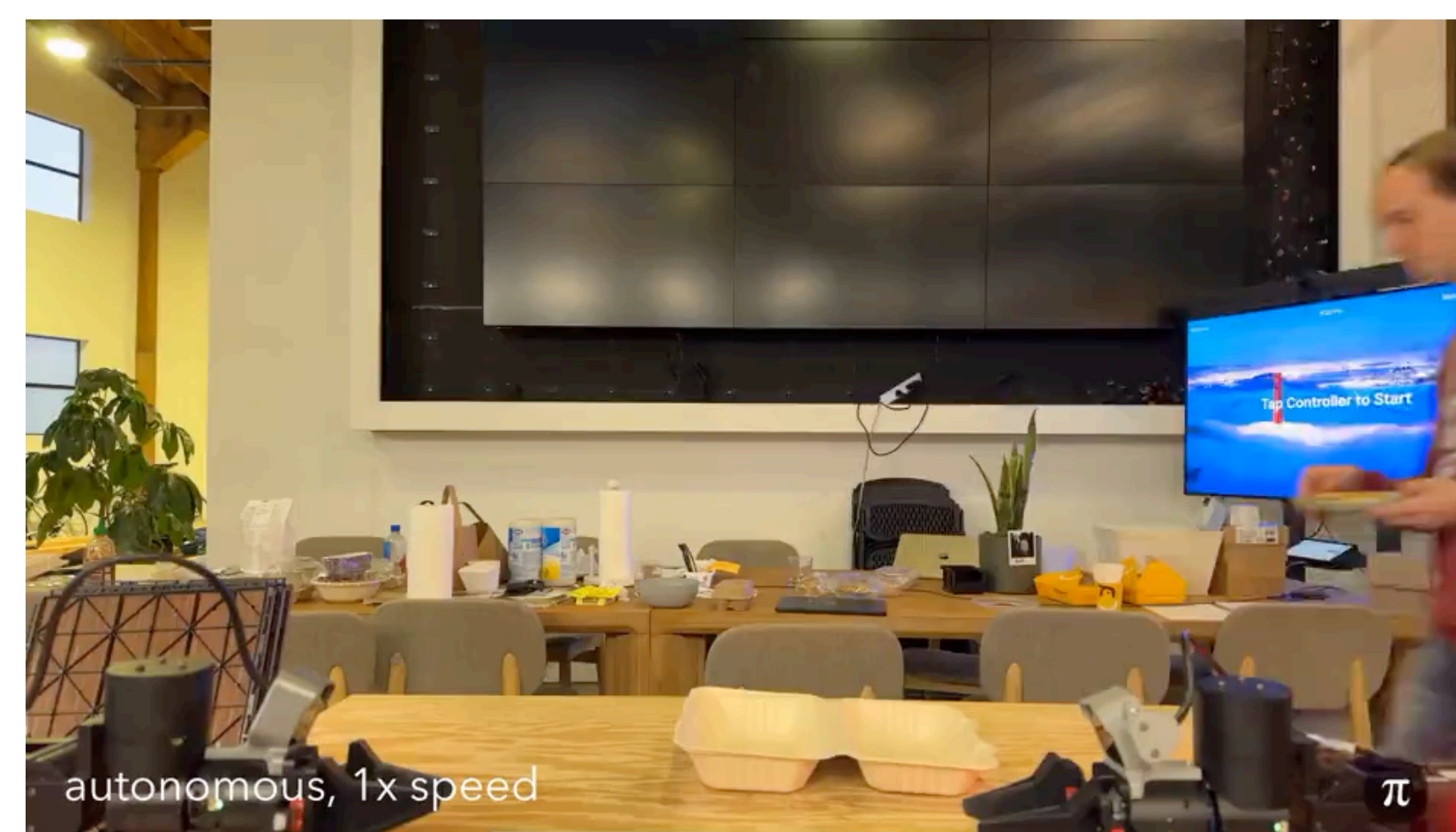
Credits: <https://www.pi.website/blog/pi0>

SEINAR · Octobre 2026



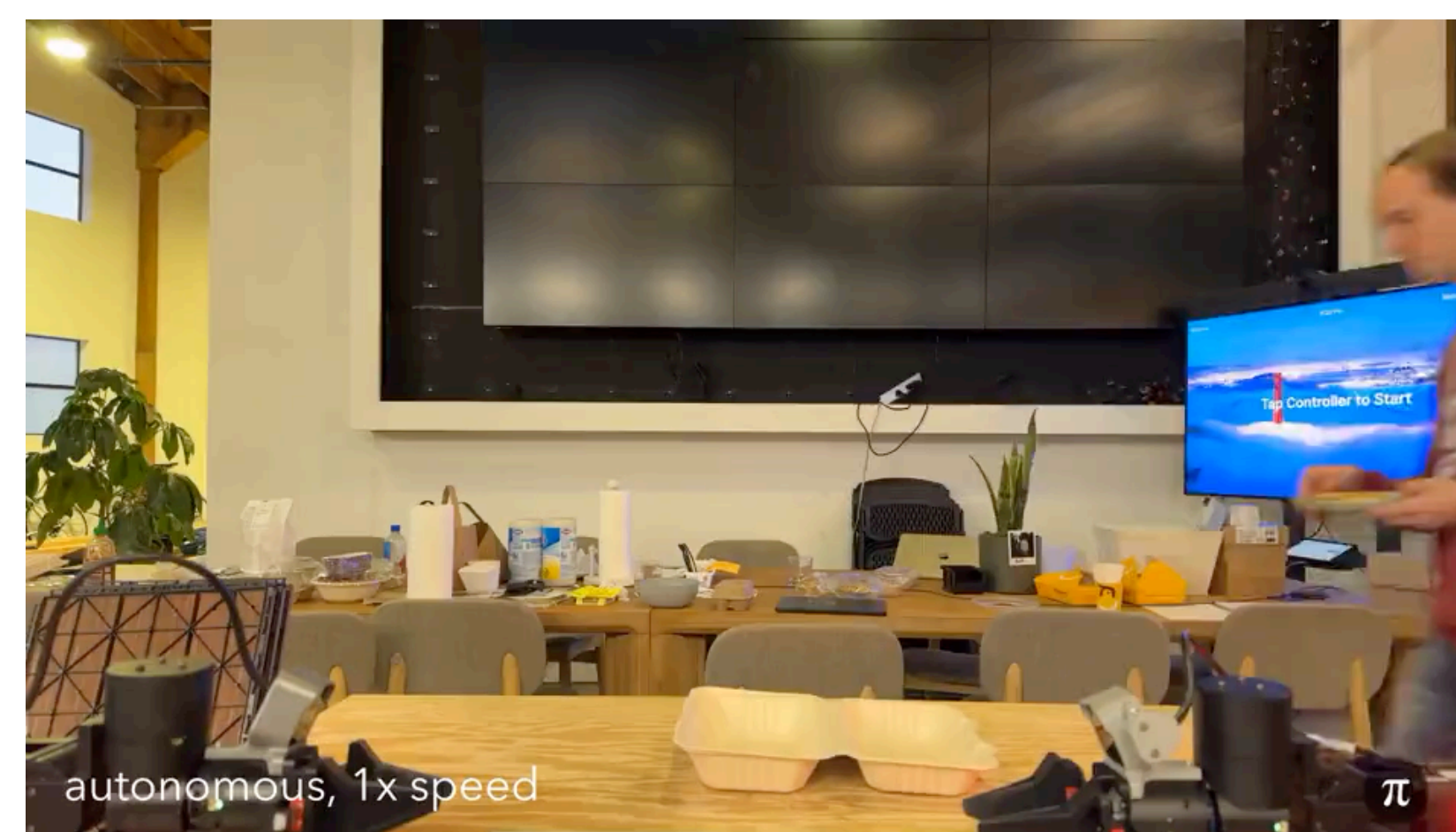
Credits: <https://www.pi.website/blog/pi0>

SEINAR · Octobre 2026



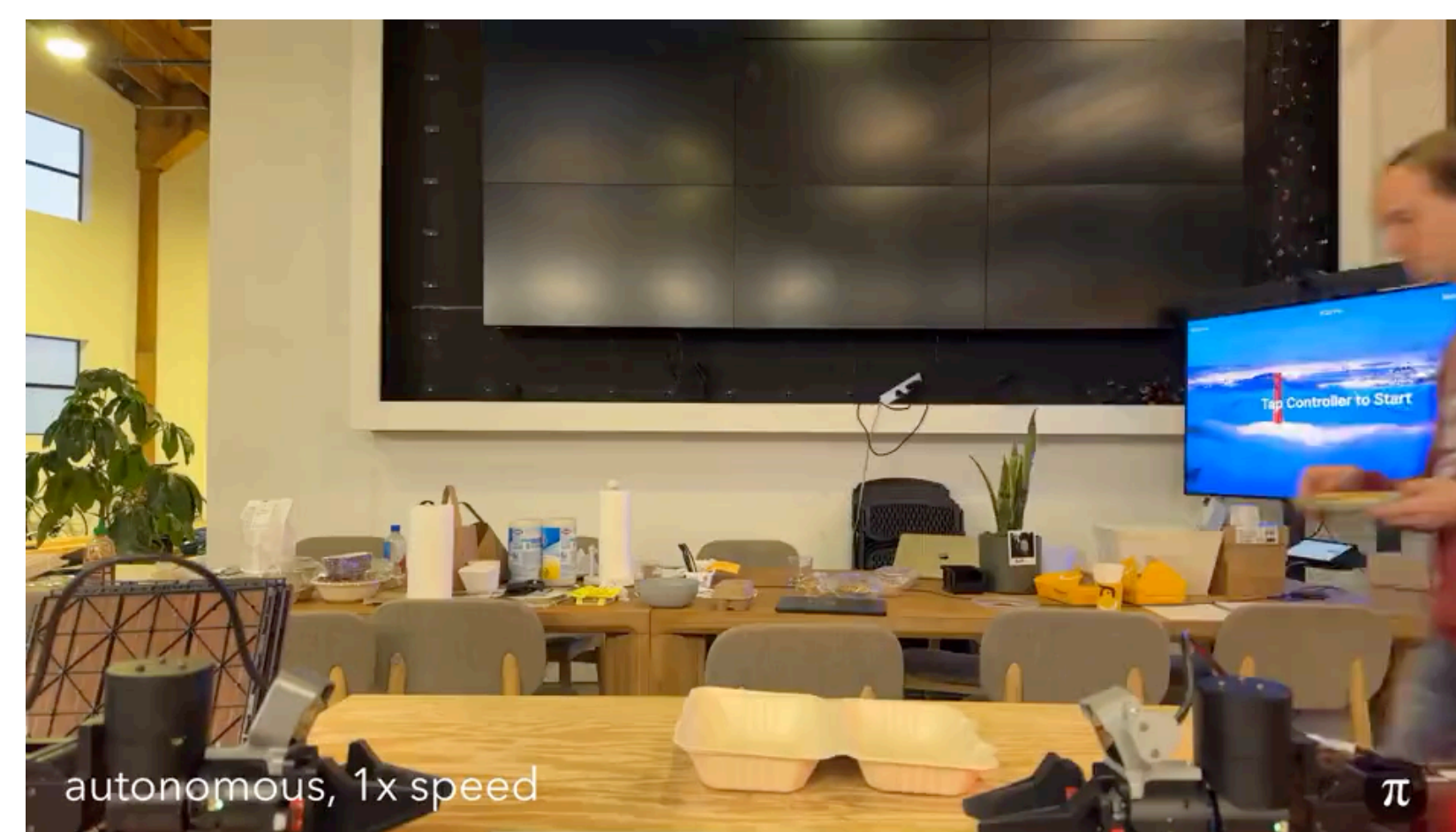
Credits: <https://www.pi.website/blog/pi0>

SEINAR · Octobre 2026



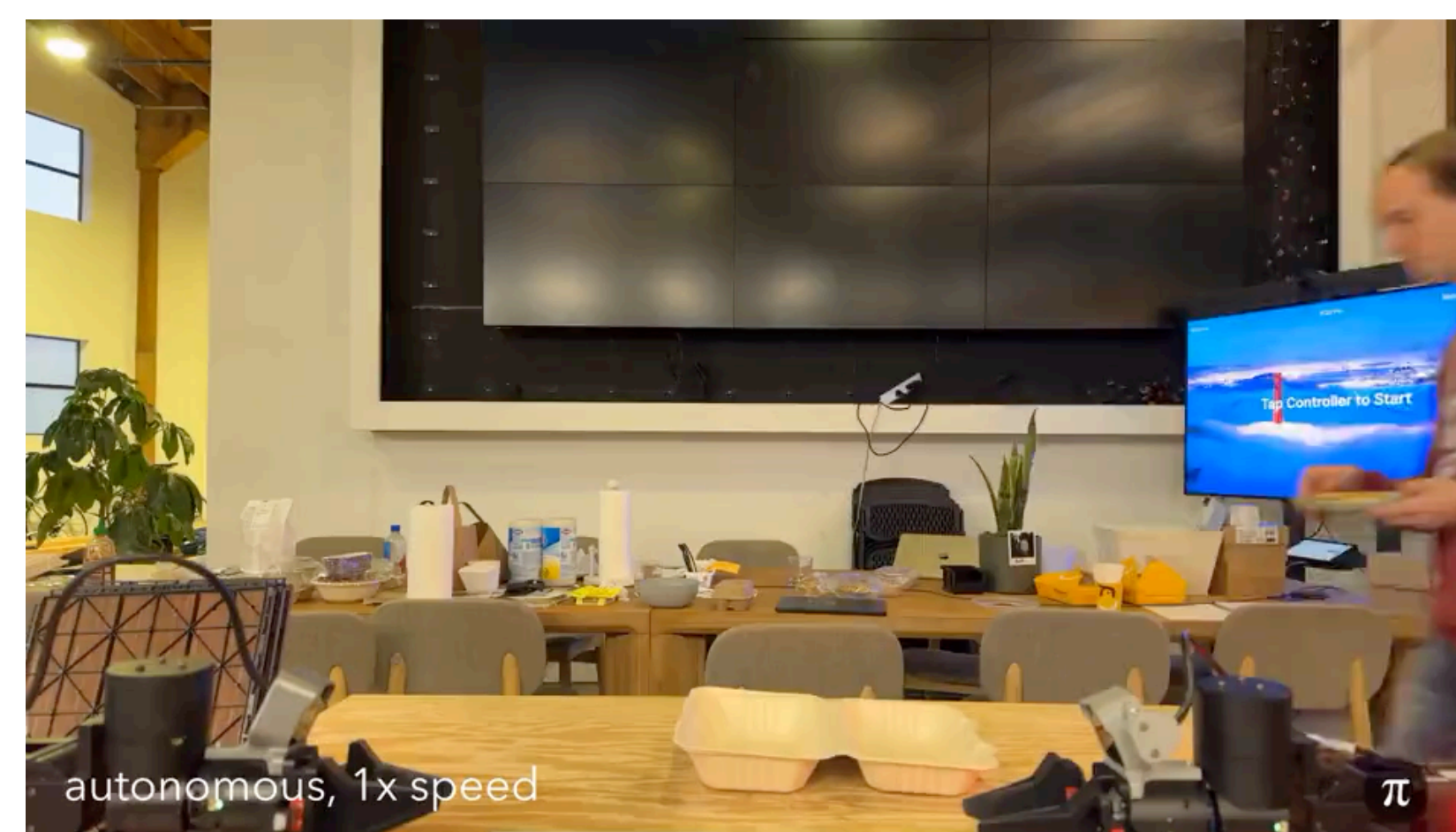
Credits: <https://www.pi.website/blog/pi0>

SEINAR · Octobre 2026



Credits: <https://www.pi.website/blog/pi0>

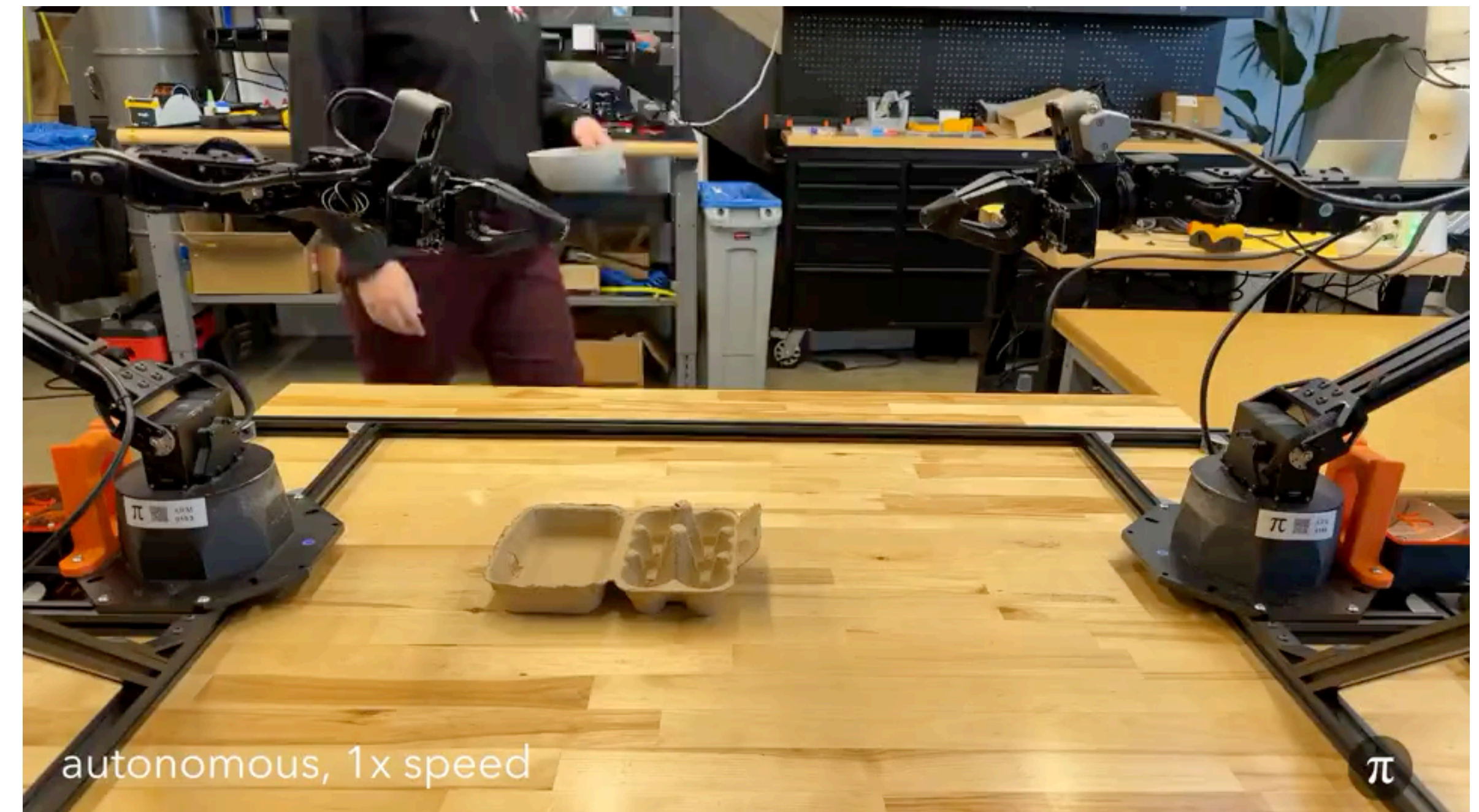
SEINAR · Octobre 2026



Credits: <https://www.pi.website/blog/pi0>

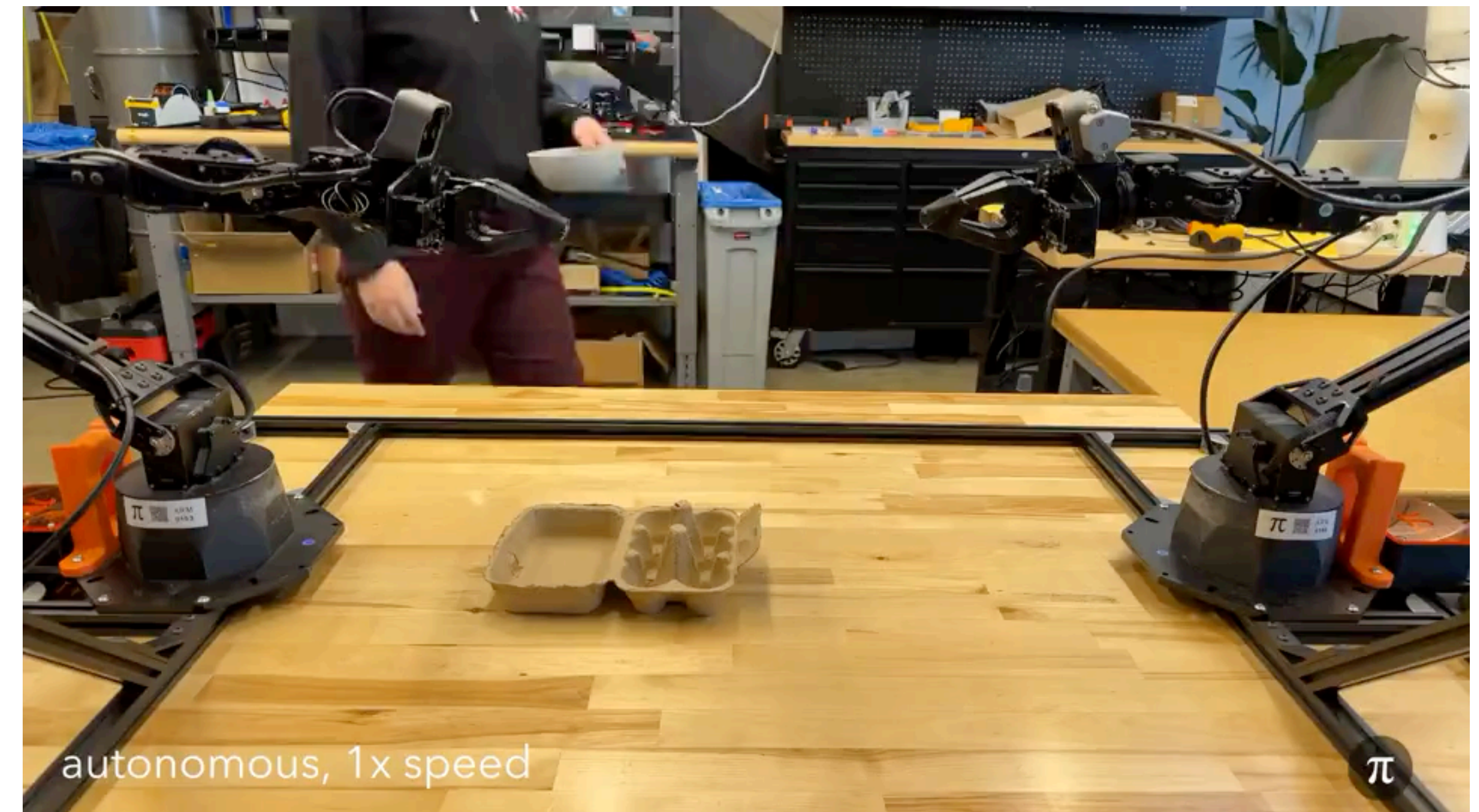
SEINAR · Octobre 2026

Une consigne, plusieurs decisions



π_0 en action

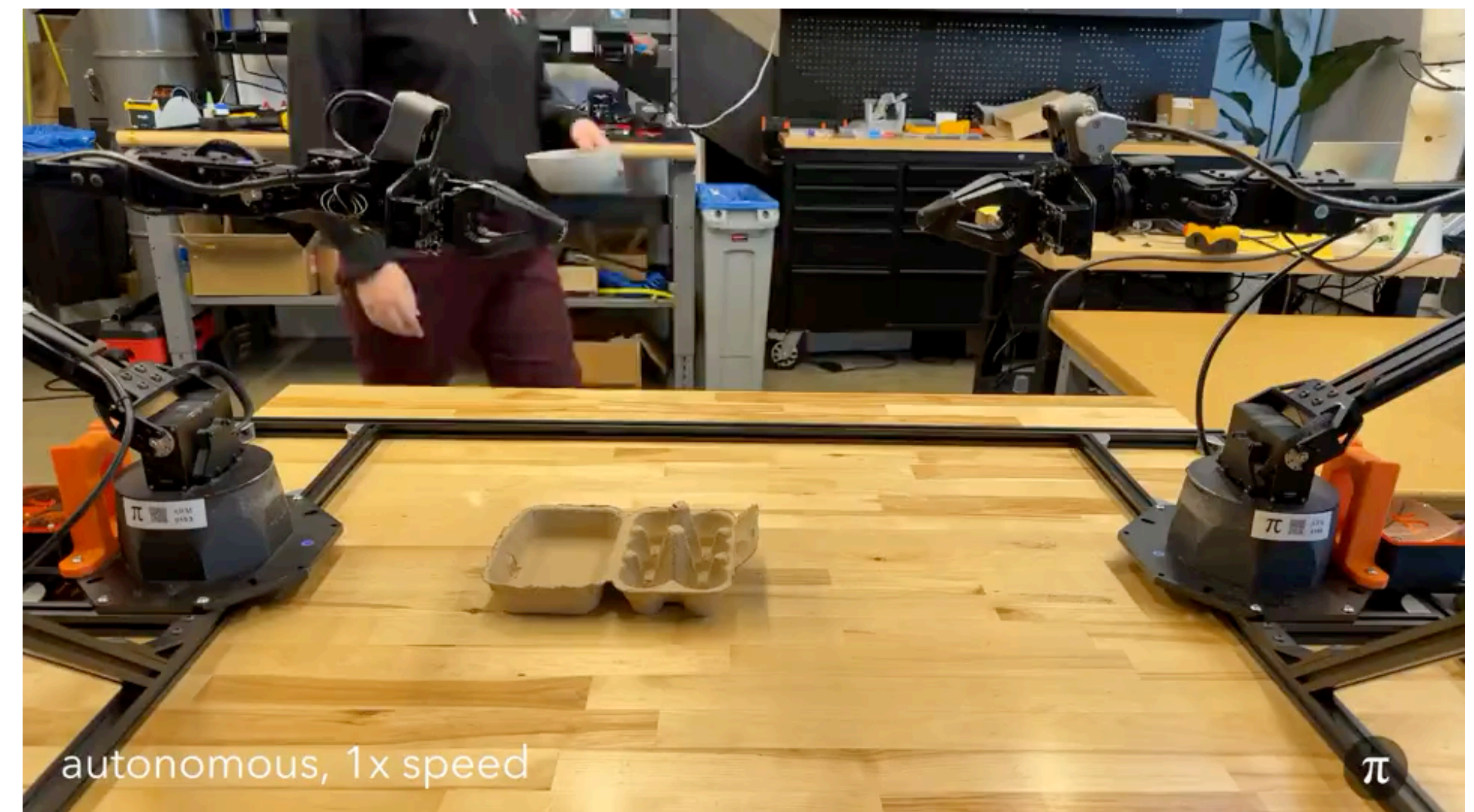
Une consigne, plusieurs decisions



π_0 en action

Une consigne, plusieurs decisions

<< Mets les œufs dans la plaquette. >>

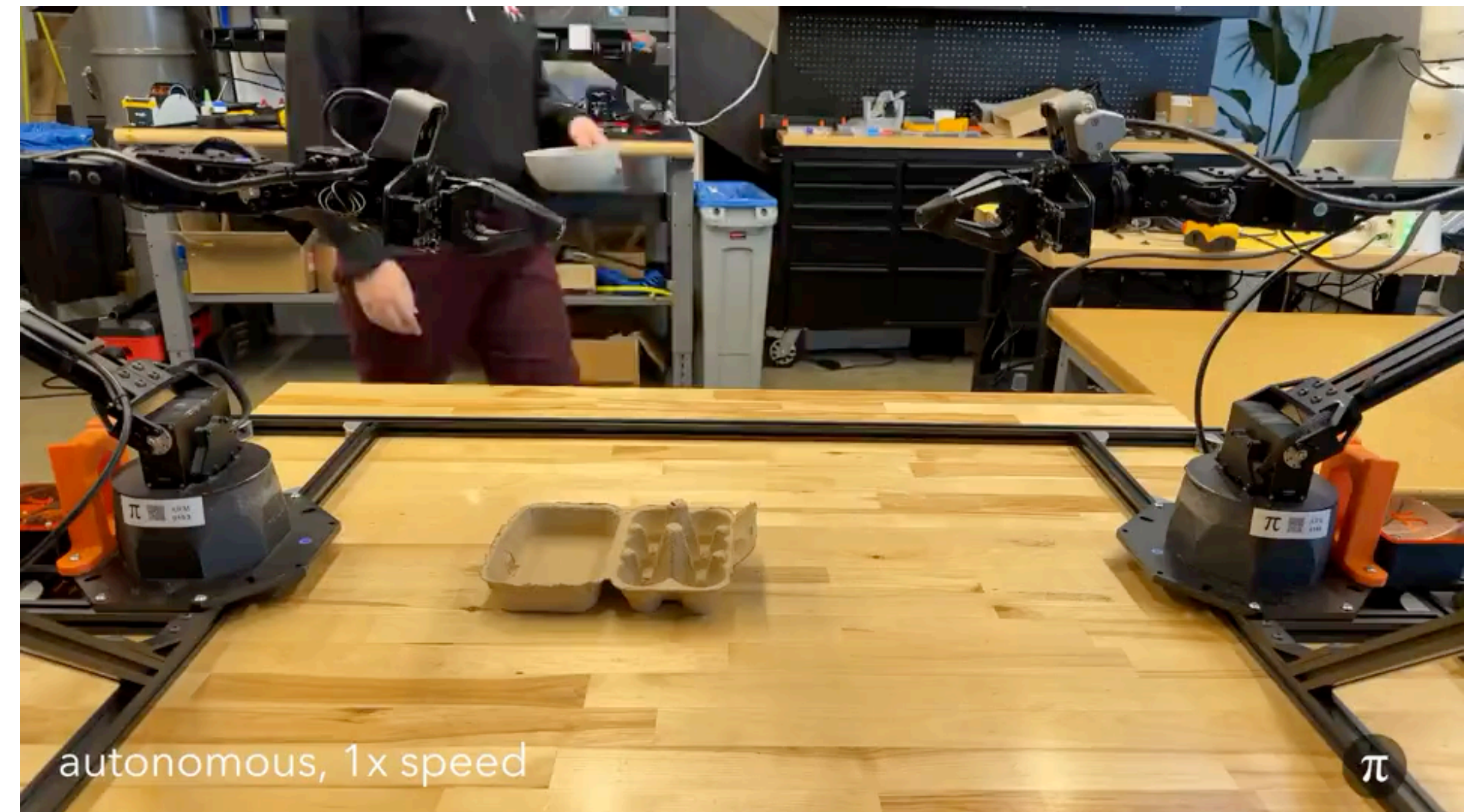


π_0 en action

Une consigne, plusieurs decisions

<< Mets les œufs dans la plaquette. >>

Quels oeufs?



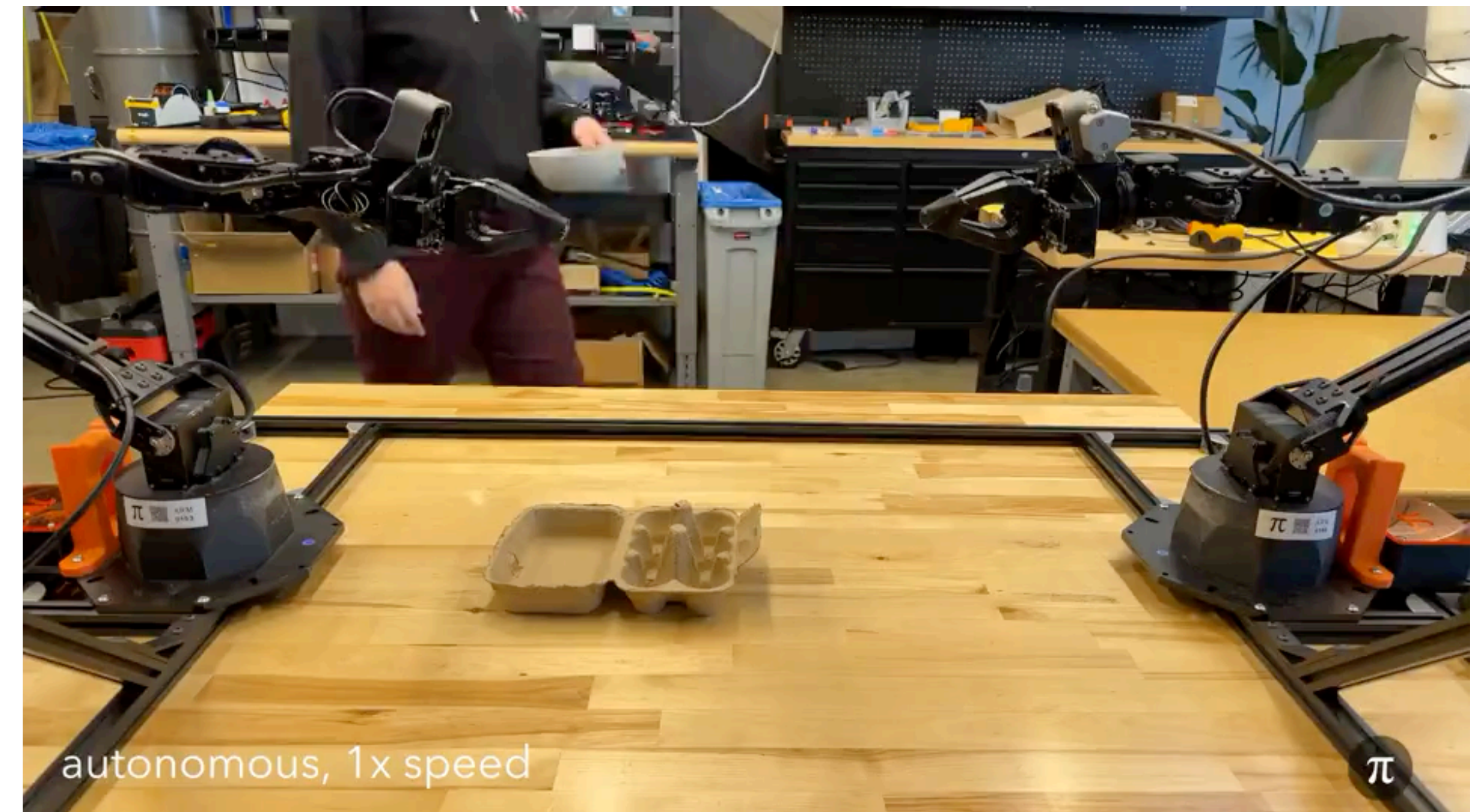
π_0 en action

Une consigne, plusieurs decisions

<< Mets les œufs dans la plaquette. >>

Quels oeufs?

Relier les mots aux
objets visibles



π_0 en action

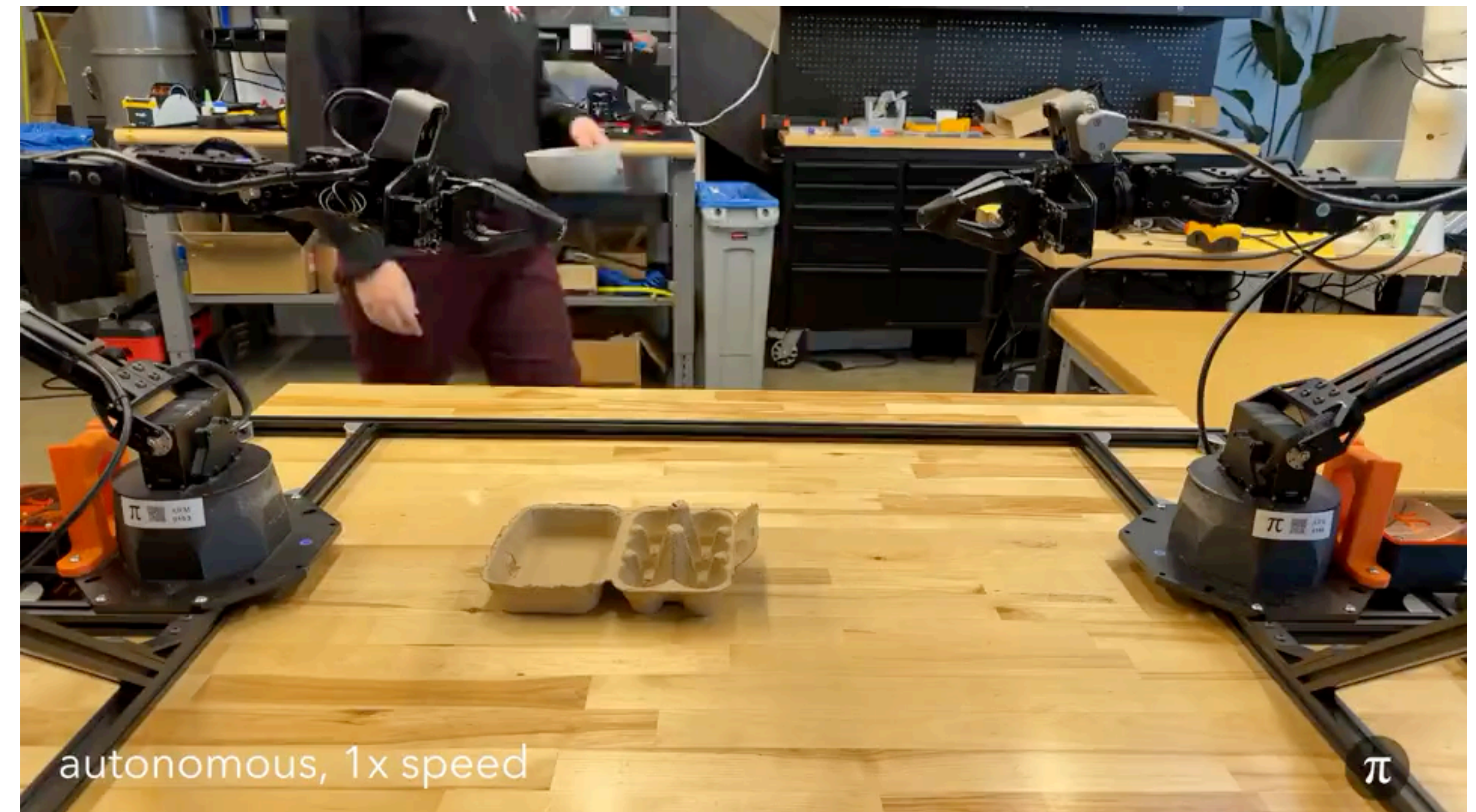
Une consigne, plusieurs decisions

<< Mets les œufs dans la plaquette. >>

Quels oeufs?

Relier les mots aux
objets visibles

Quels mouvements?



π_0 en action

Une consigne, plusieurs decisions

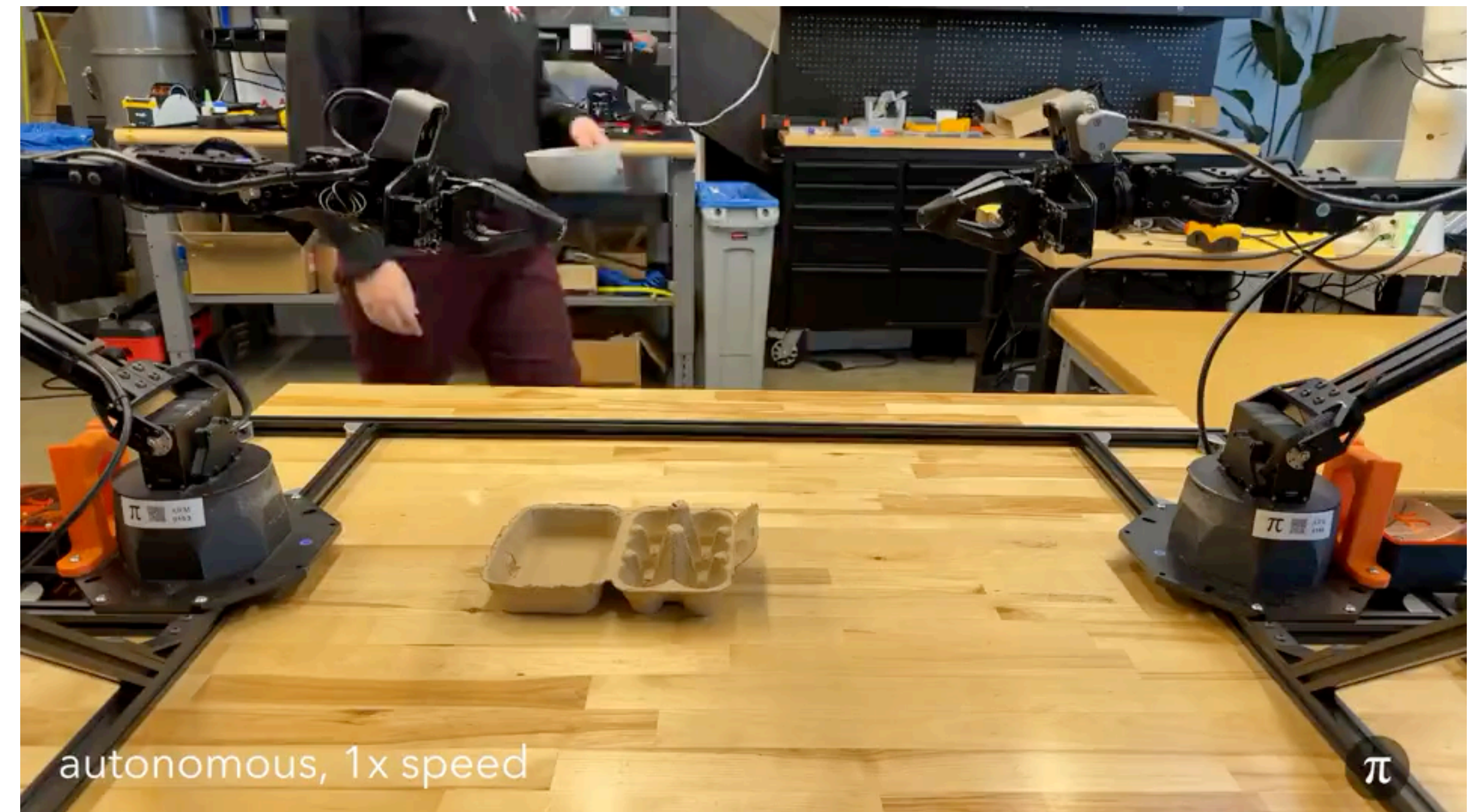
<< Mets les œufs dans la plaquette. >>

Quels oeufs?

Relier les mots aux
objets visibles

Quels mouvements?

Choisir un prise et une
séquence d'action



π_0 en action

Une consigne, plusieurs decisions

<< Mets les œufs dans la plaquette. >>

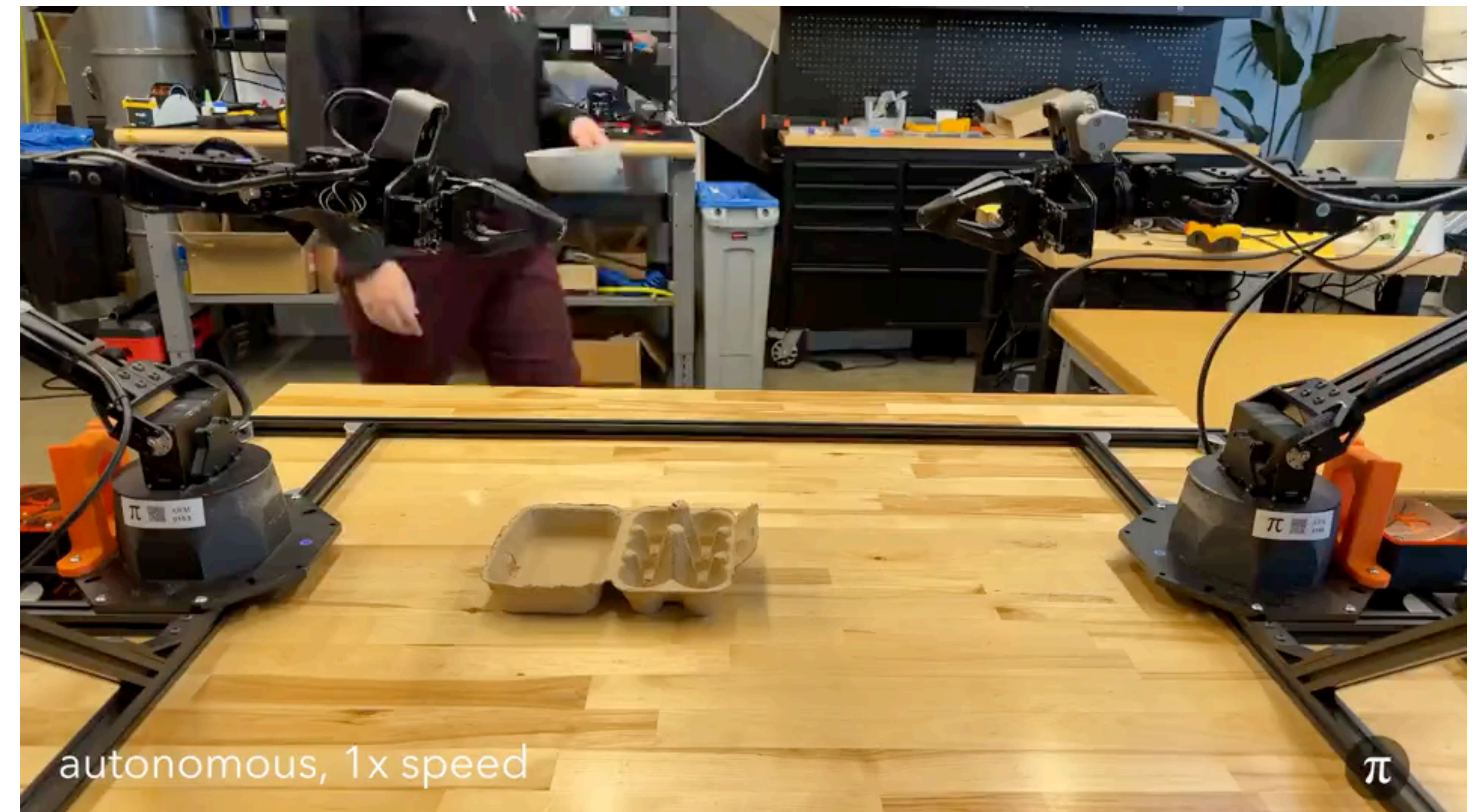
Quels oeufs?

Relier les mots aux
objets visibles

Quels mouvements?

Choisir un prise et une
séquence d'action

La tâche est-elle
finie?



π_0 en action

Une consigne, plusieurs decisions

<< Mets les œufs dans la plaquette. >>

Quels oeufs?

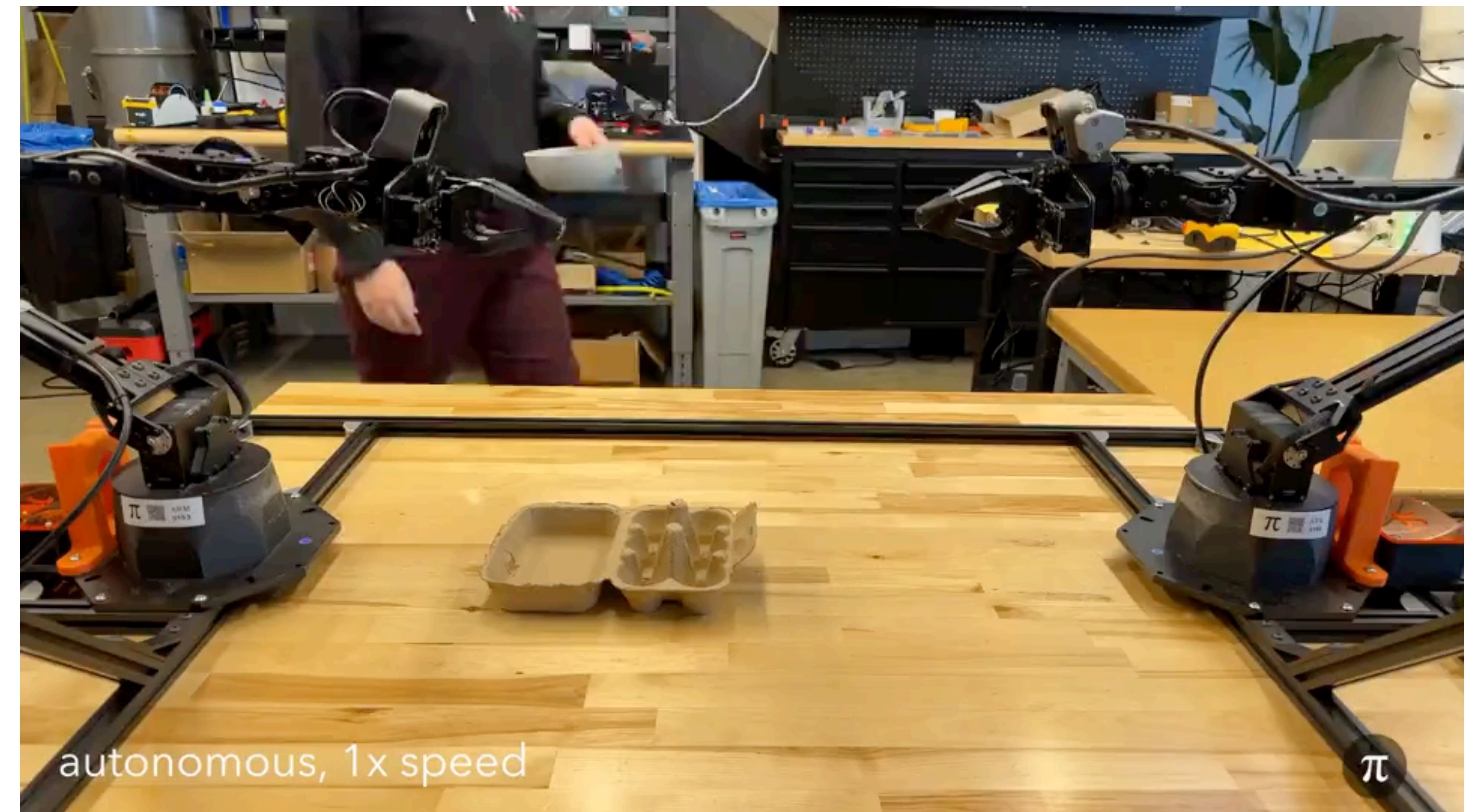
Relier les mots aux
objets visibles

Quels mouvements?

Choisir un prise et une
séquence d'action

La tâche est-elle
finie?

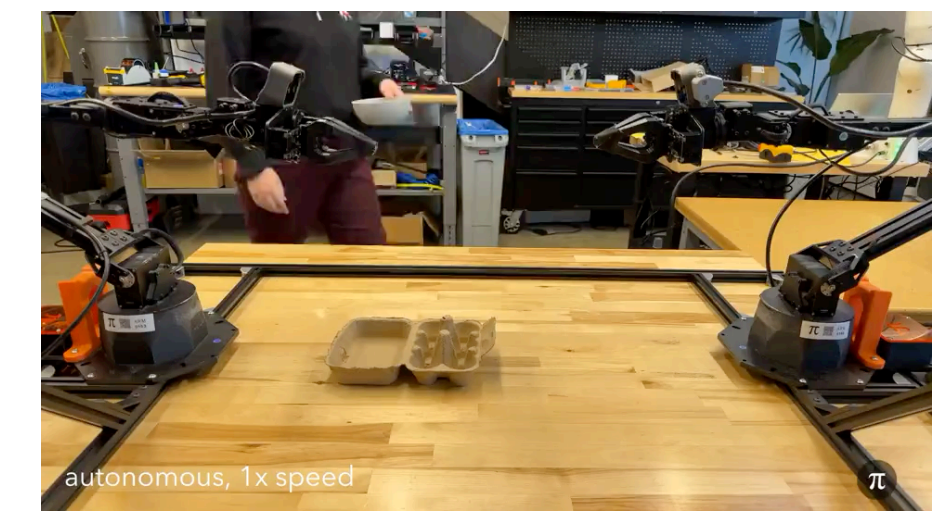
Observer le resultat et
corriger le comportement



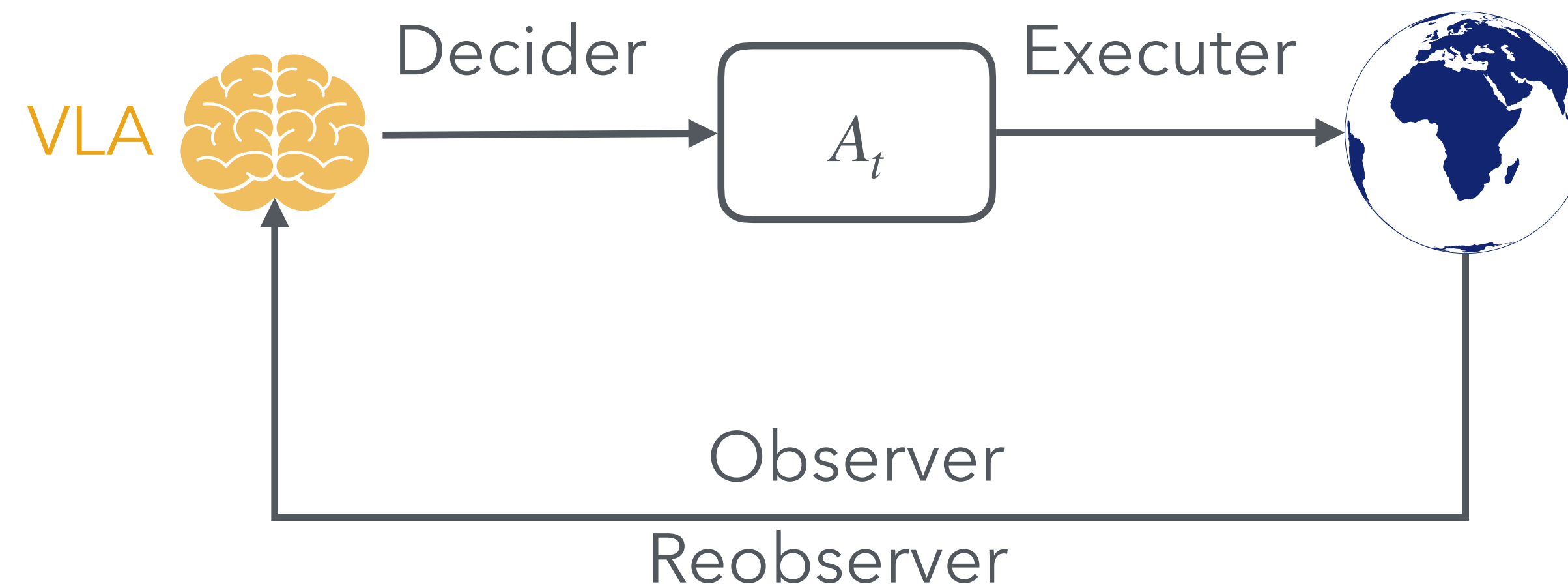
π_0 en action

01 Definition

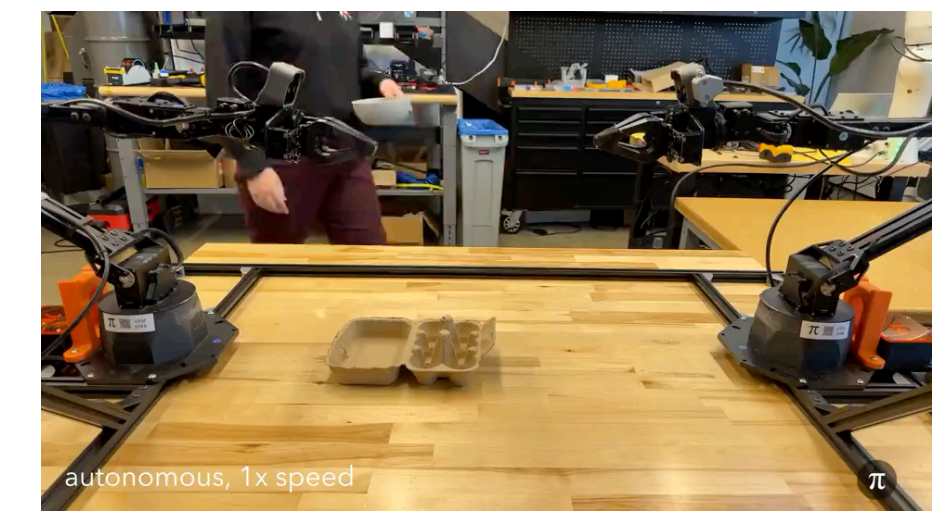
La politique dans la boucle du robot



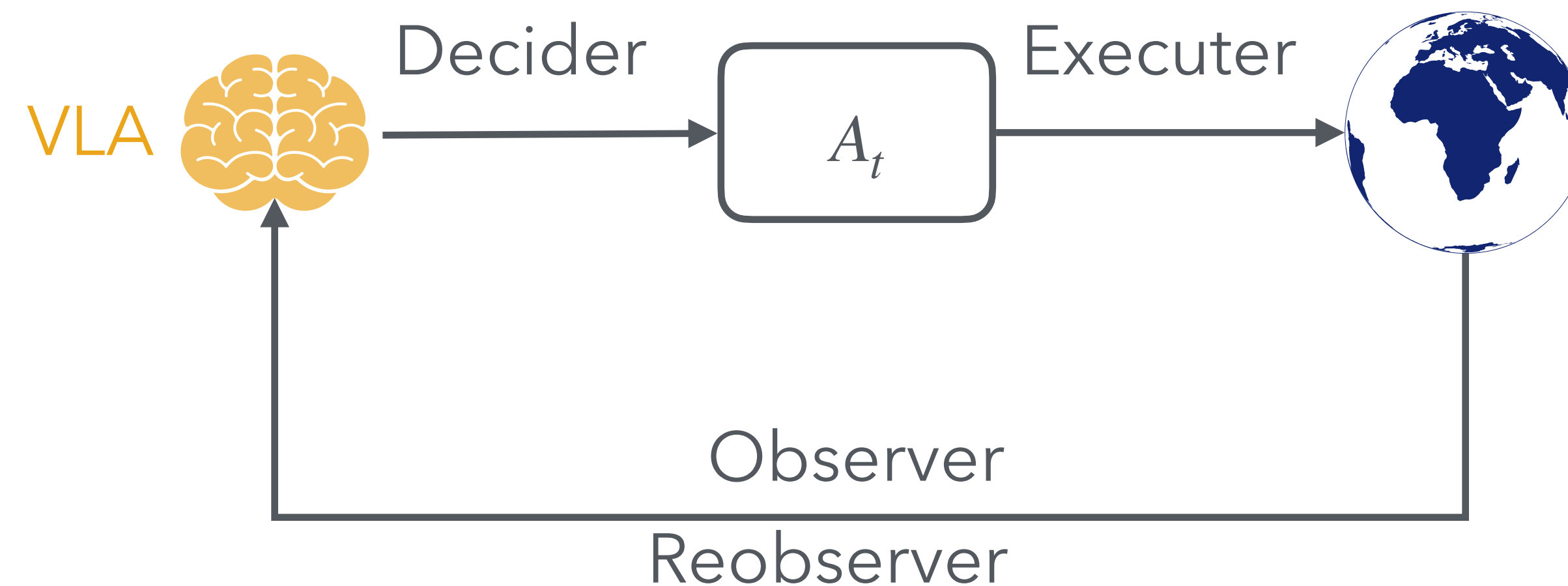
Observer → Decider → Executer → Reobserver



La politique dans la boucle du robot



Observer → Decider → Executer → Reobserver



OpenVLA: Open-source Vision-Language-Action model

LLM

Modèle de langage de base (backbone)

ViT

Modèle de Vision (Vision Transformer)

OpenVLA: Open-source Vision-Language-Action model



Input image

LLM

ViT



OpenVLA: Open-source Vision-Language-Action model



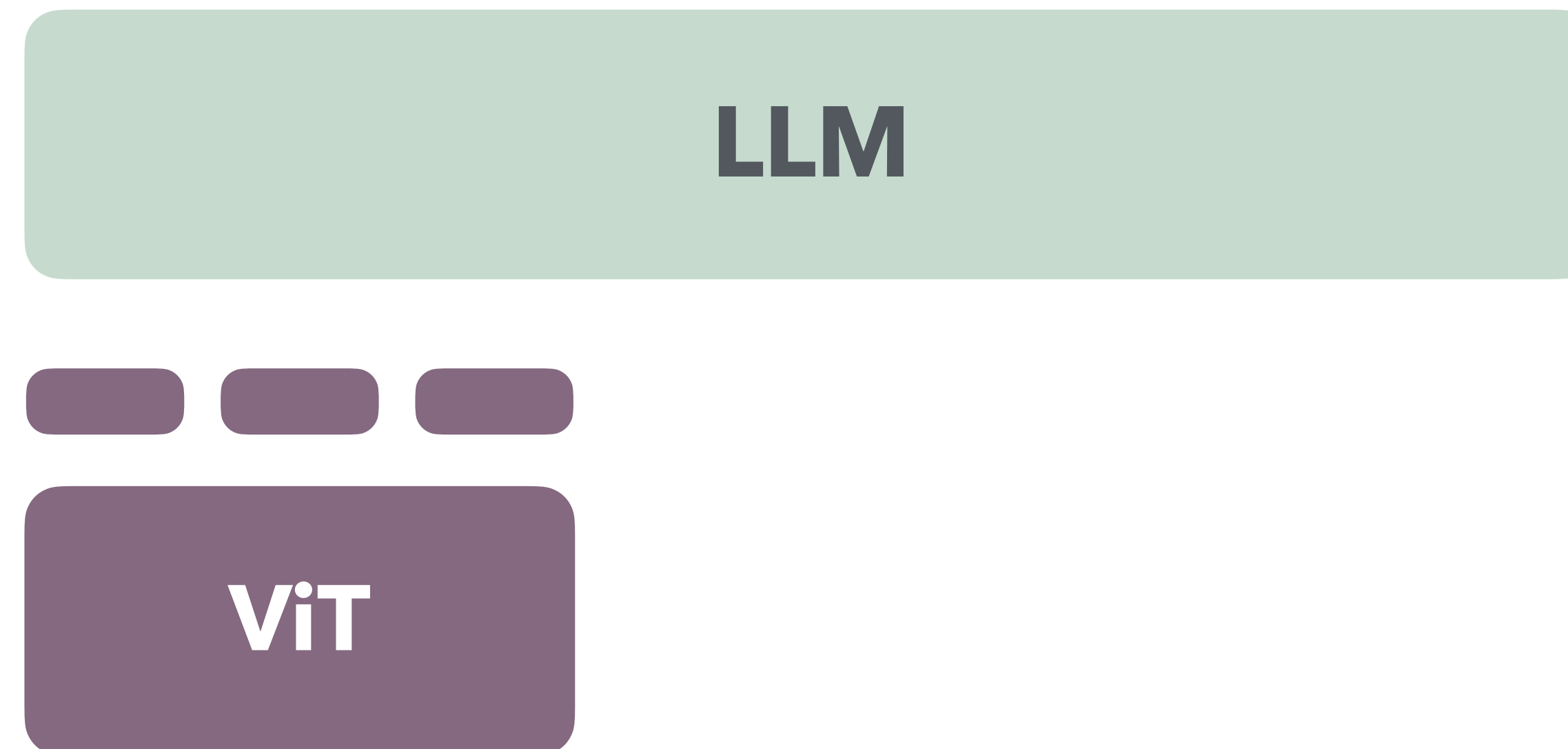
Input image



OpenVLA: Open-source Vision-Language-Action model



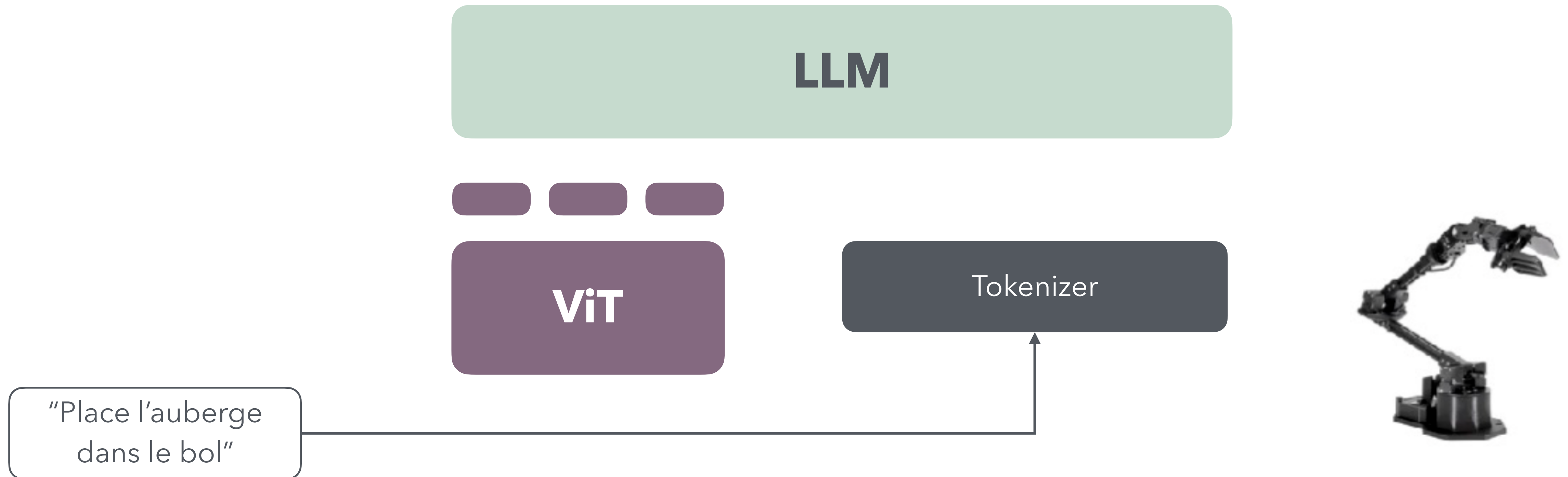
OpenVLA: Open-source Vision-Language-Action model



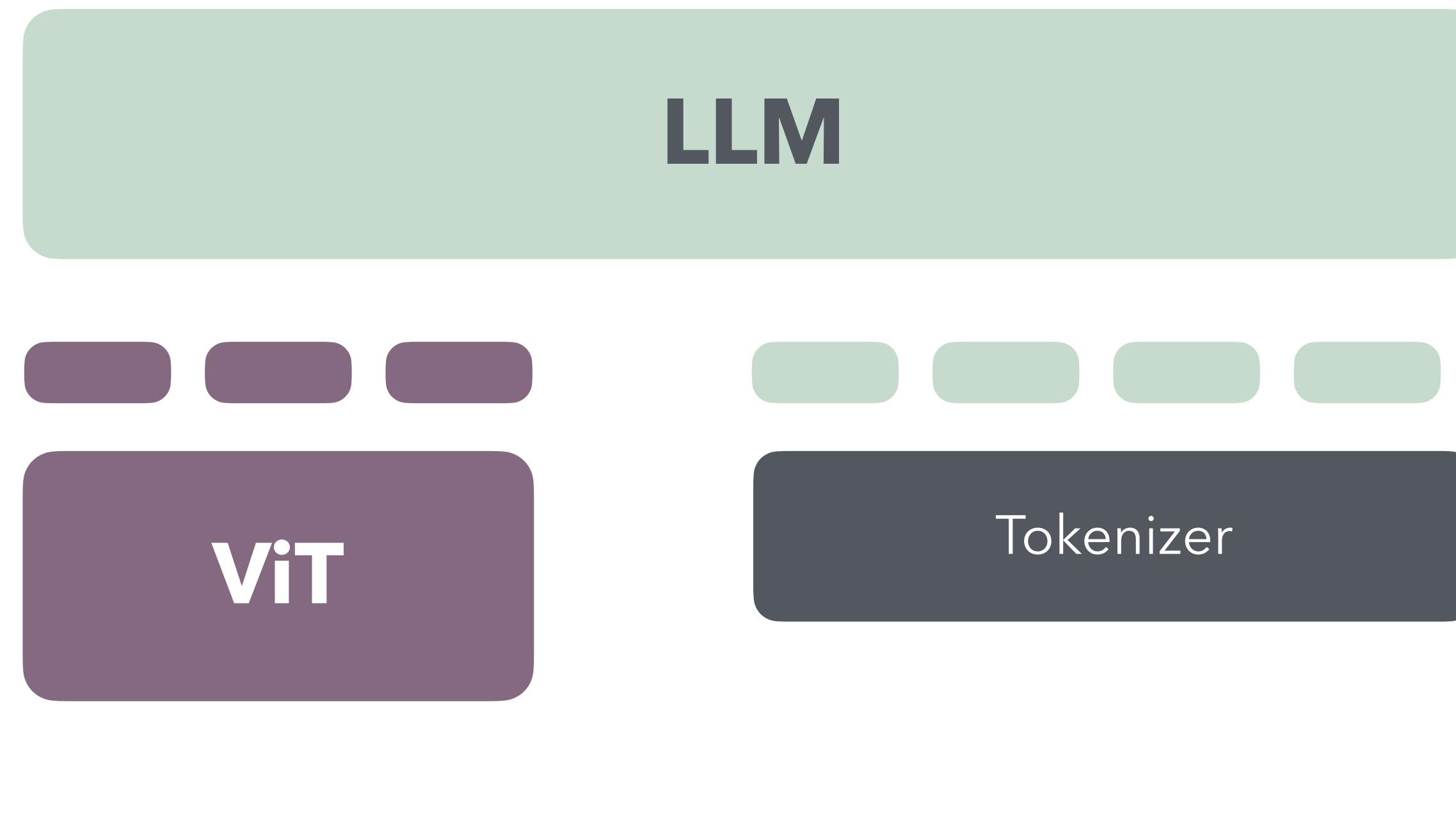
"Place l'auberge
dans le bol"



OpenVLA: Open-source Vision-Language-Action model



OpenVLA: Open-source Vision-Language-Action model



OpenVLA: Open-source Vision-Language-Action model

LLM

ViT

Tokenizer



OpenVLA: Open-source Vision-Language-Action model

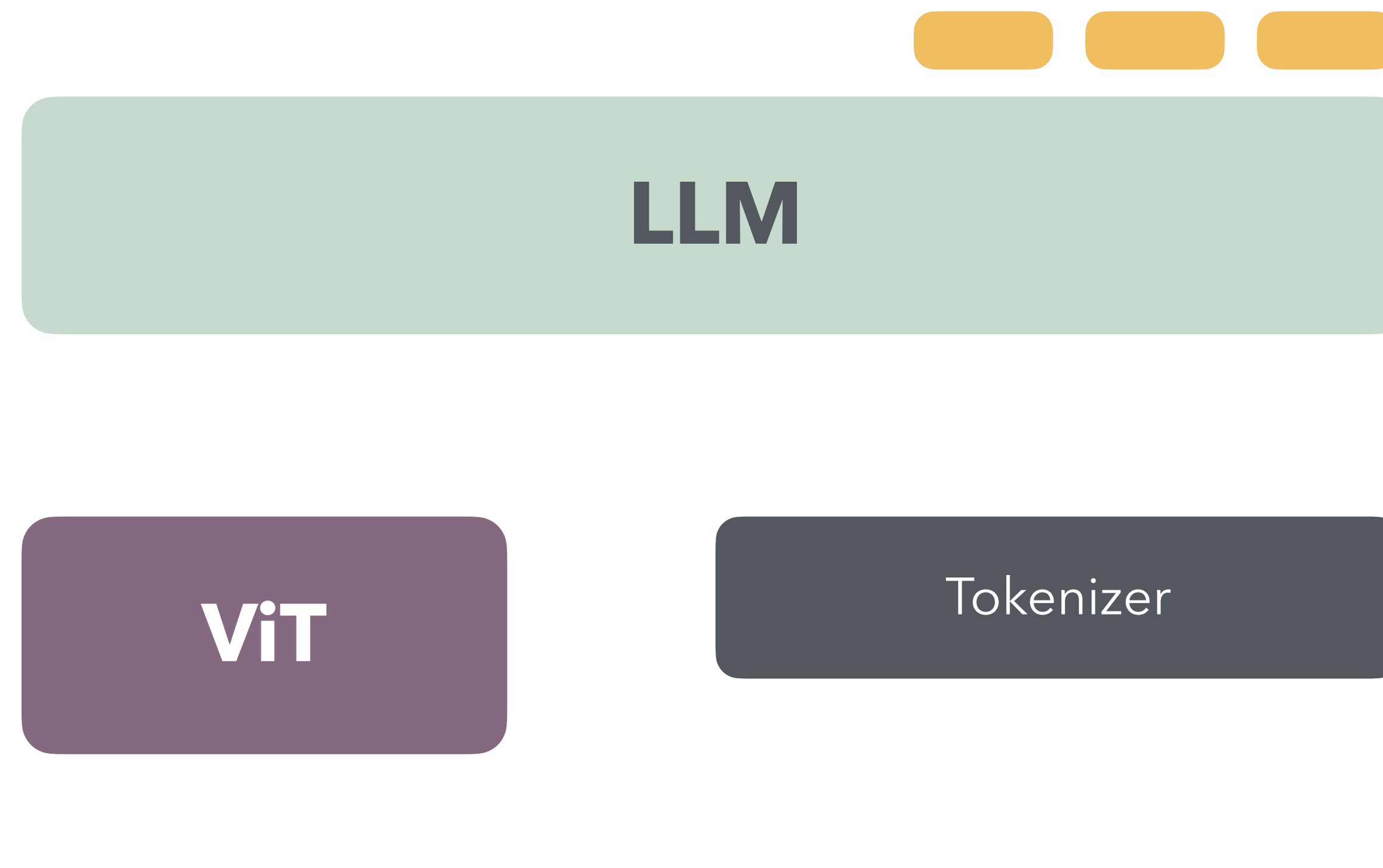
LLM

ViT

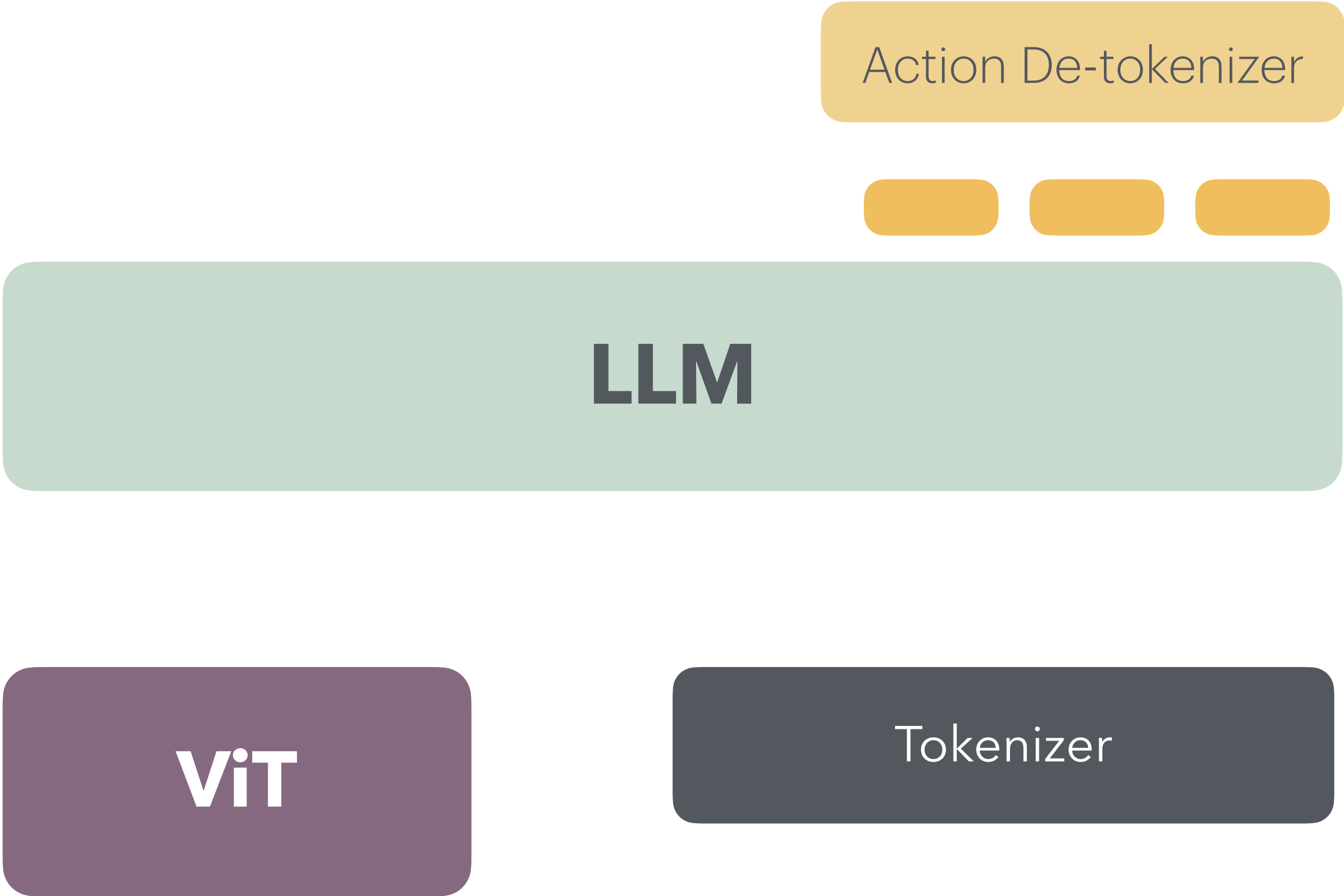
Tokenizer



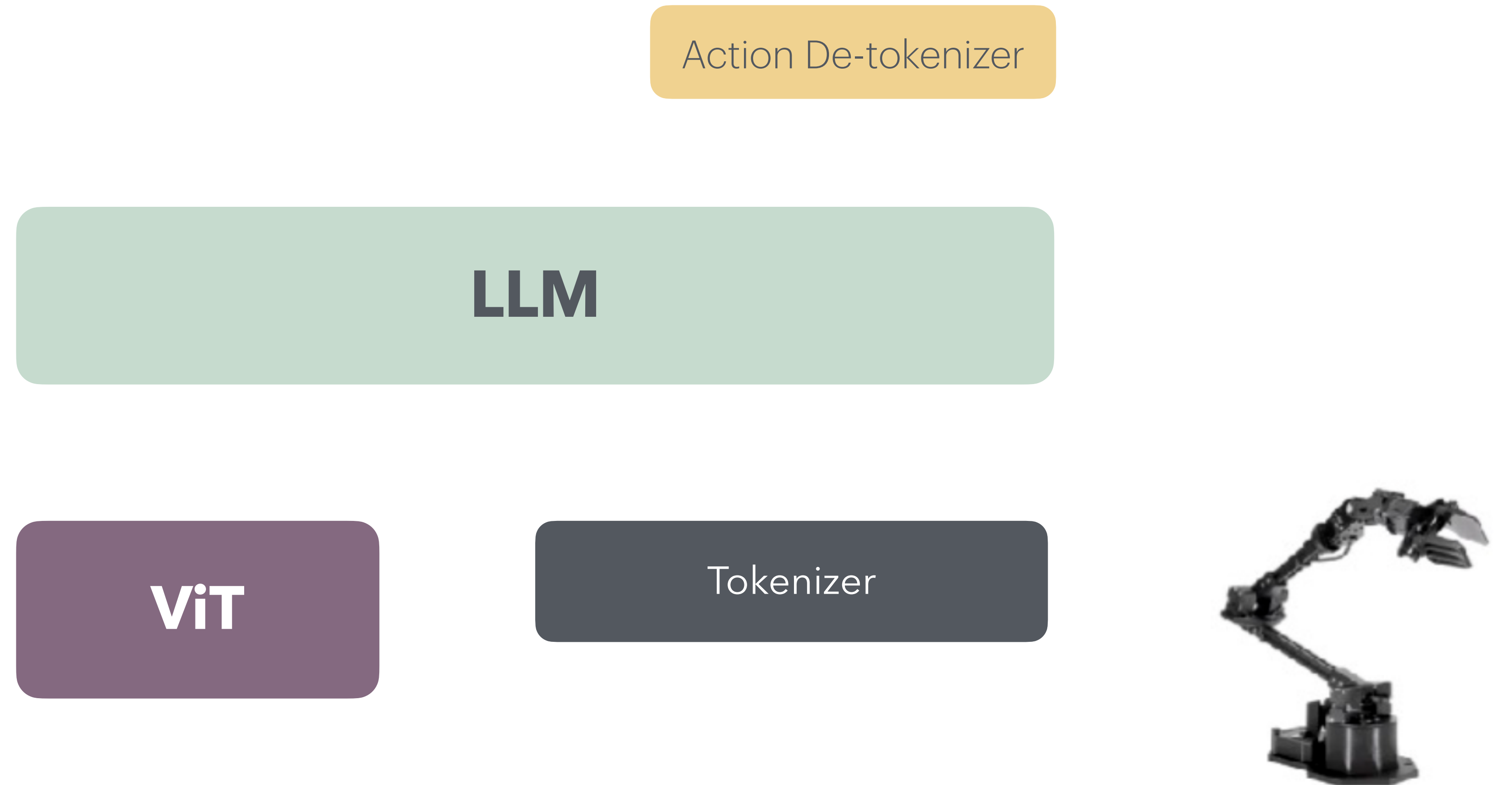
OpenVLA: Open-source Vision-Language-Action model



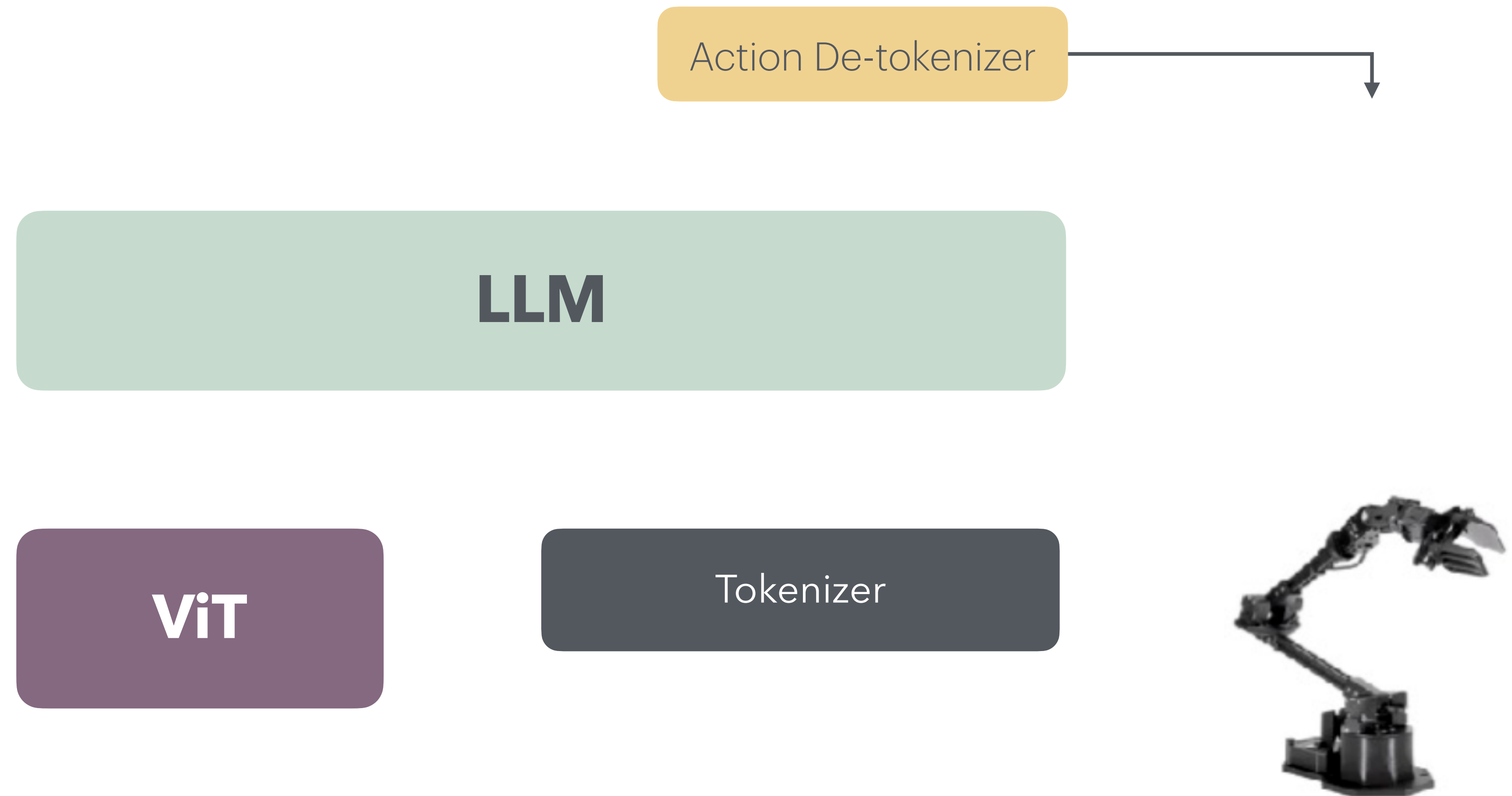
OpenVLA: Open-source Vision-Language-Action model



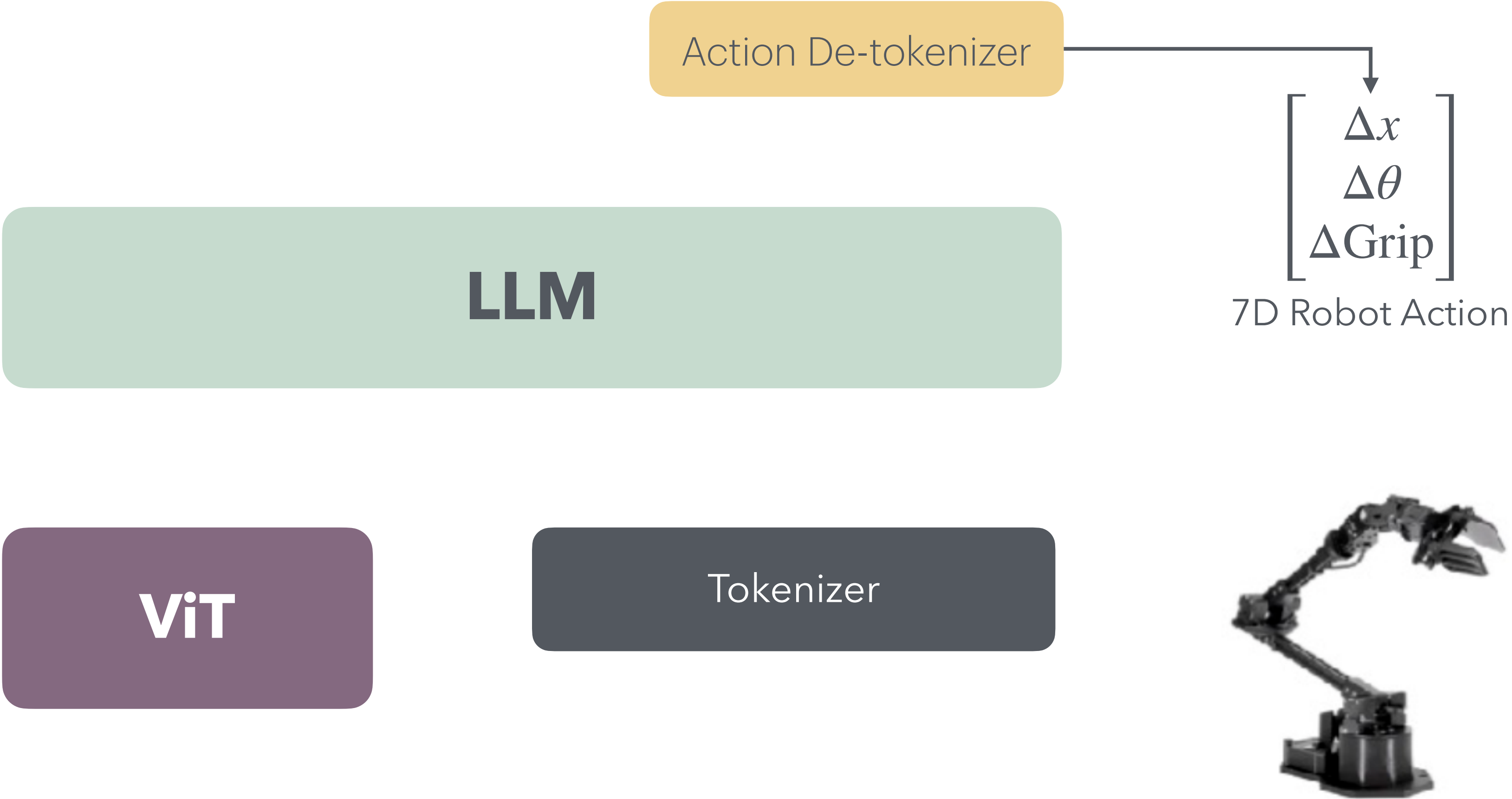
OpenVLA: Open-source Vision-Language-Action model



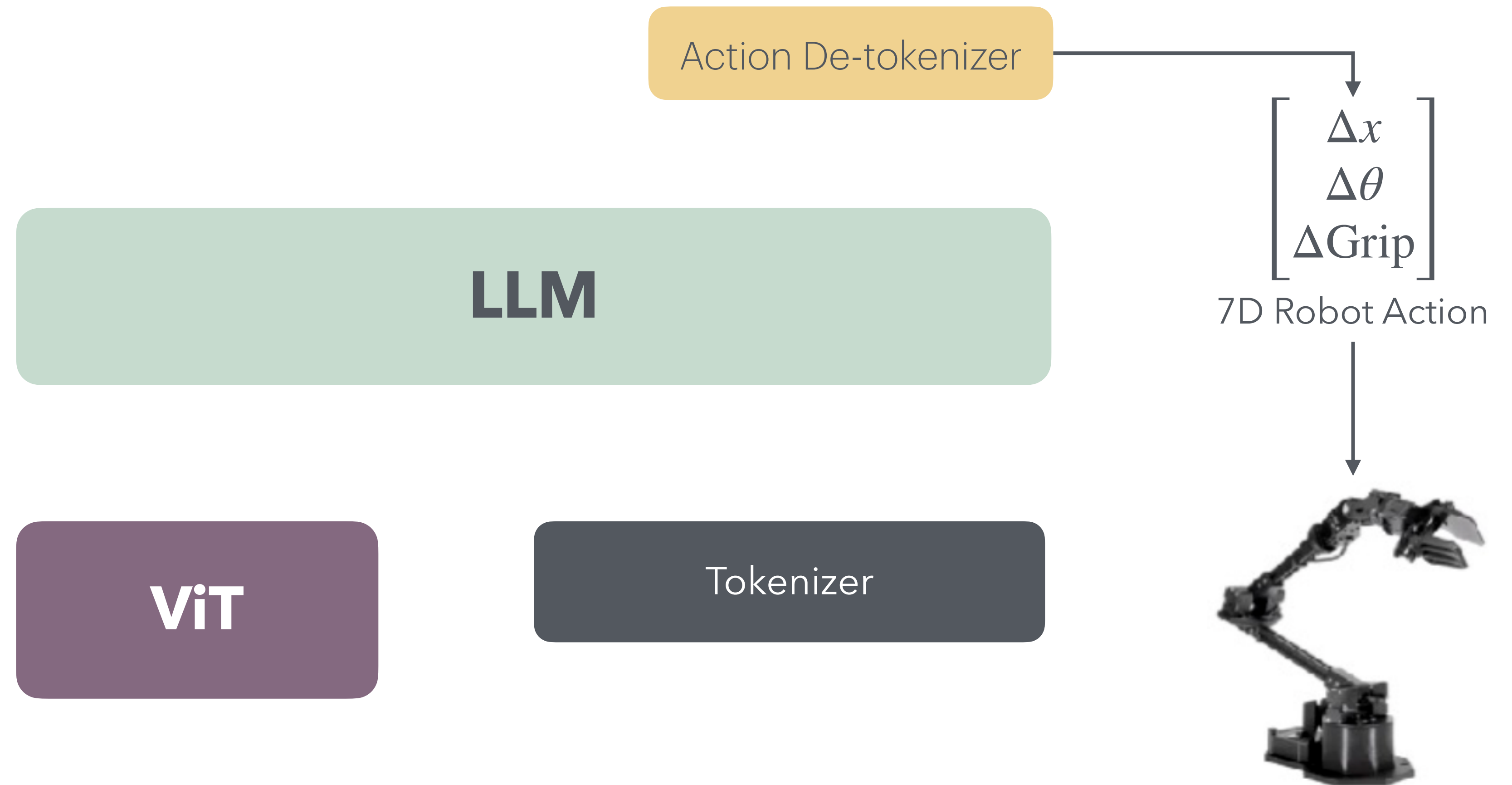
OpenVLA: Open-source Vision-Language-Action model



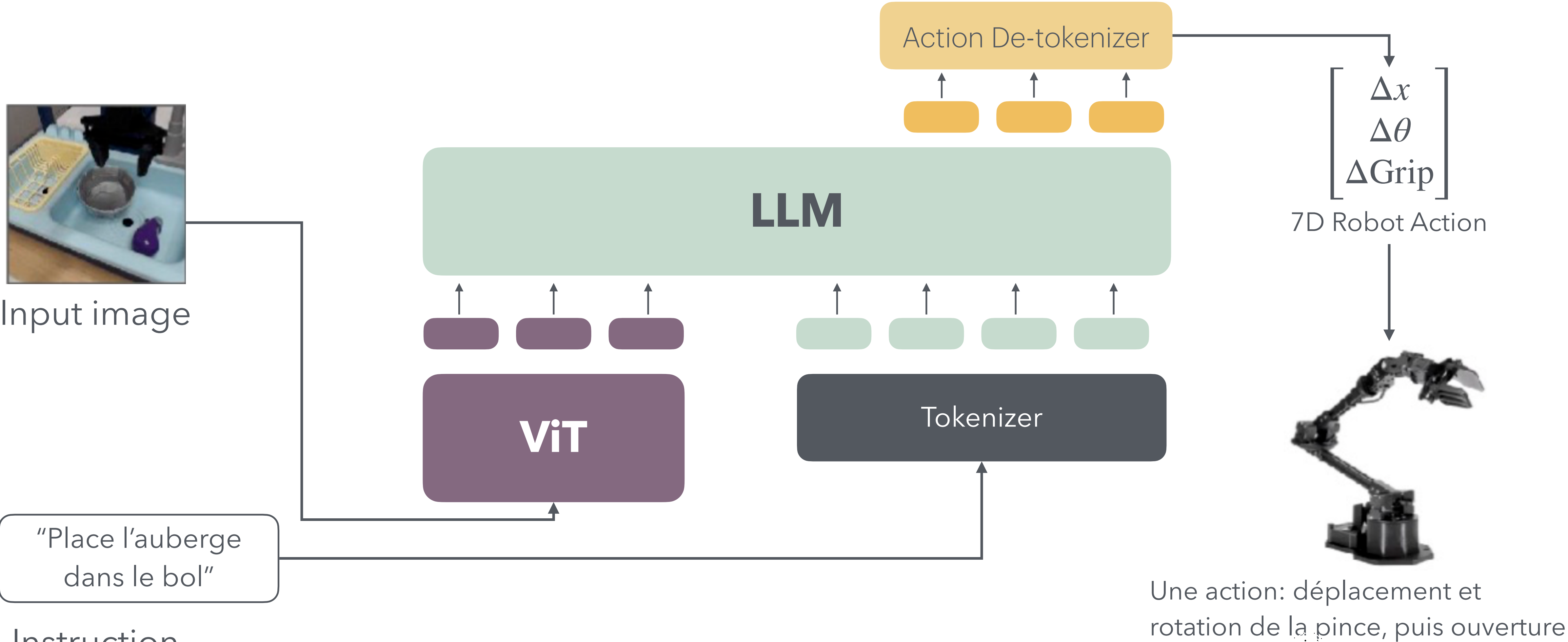
OpenVLA: Open-source Vision-Language-Action model



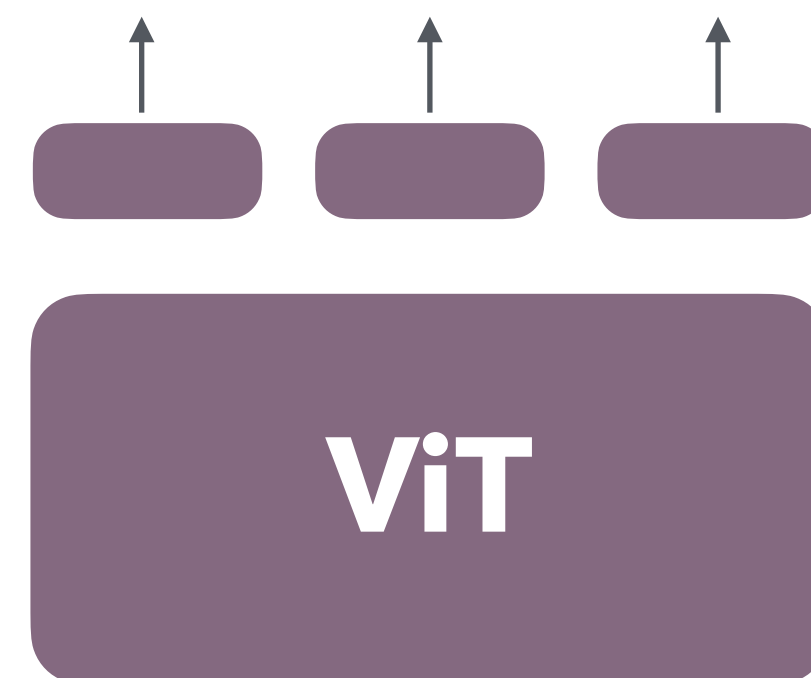
OpenVLA: Open-source Vision-Language-Action model



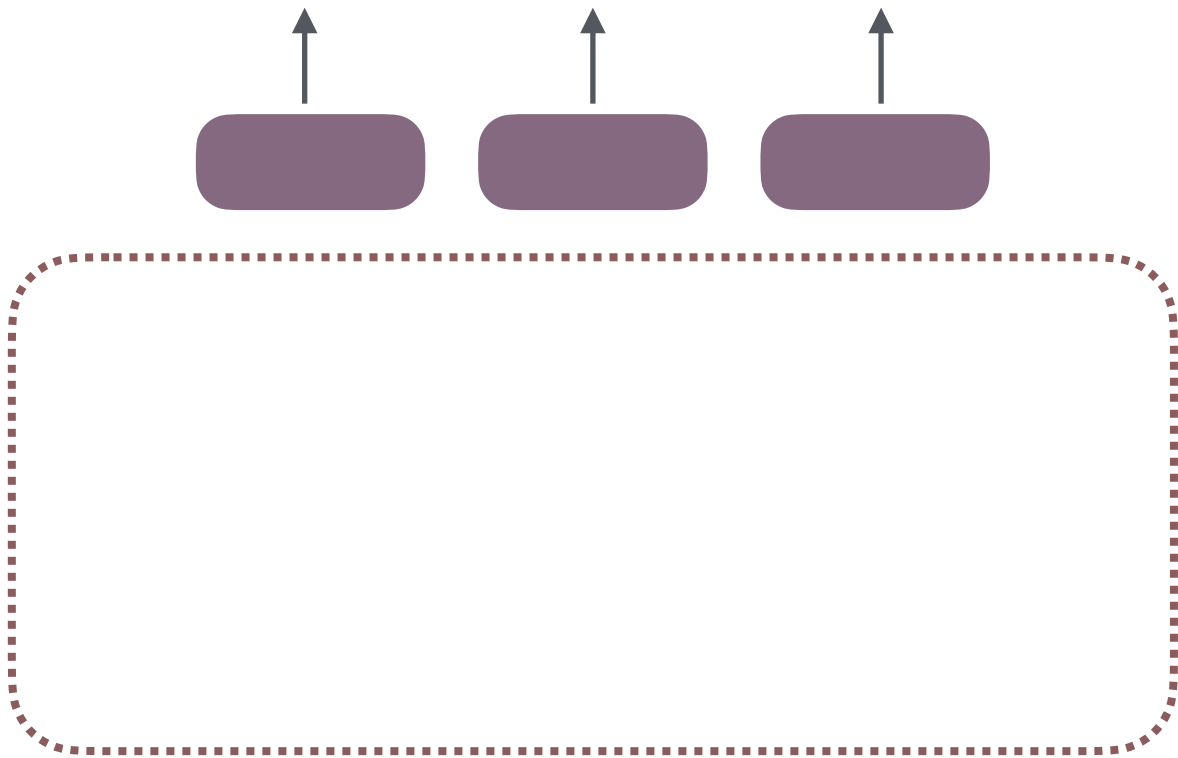
OpenVLA: Open-source Vision-Language-Action model



OpenVLA: Open-source Vision-Language-Action model

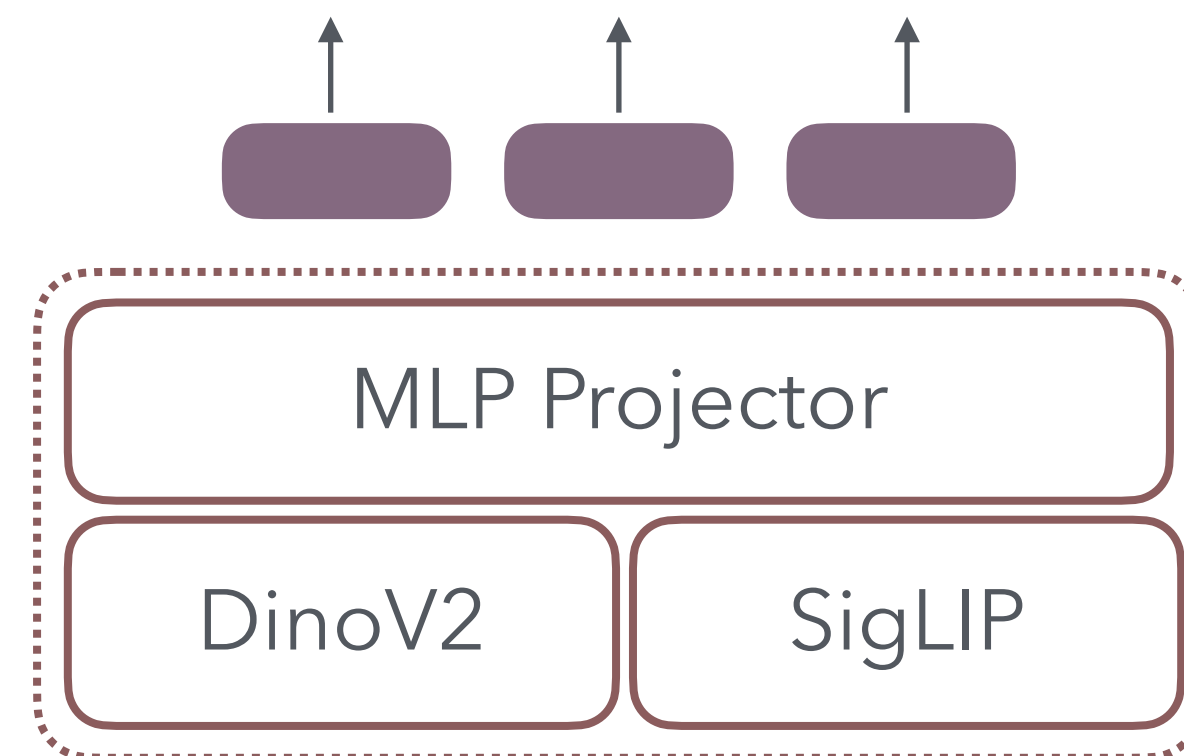


OpenVLA: Open-source Vision-Language-Action model



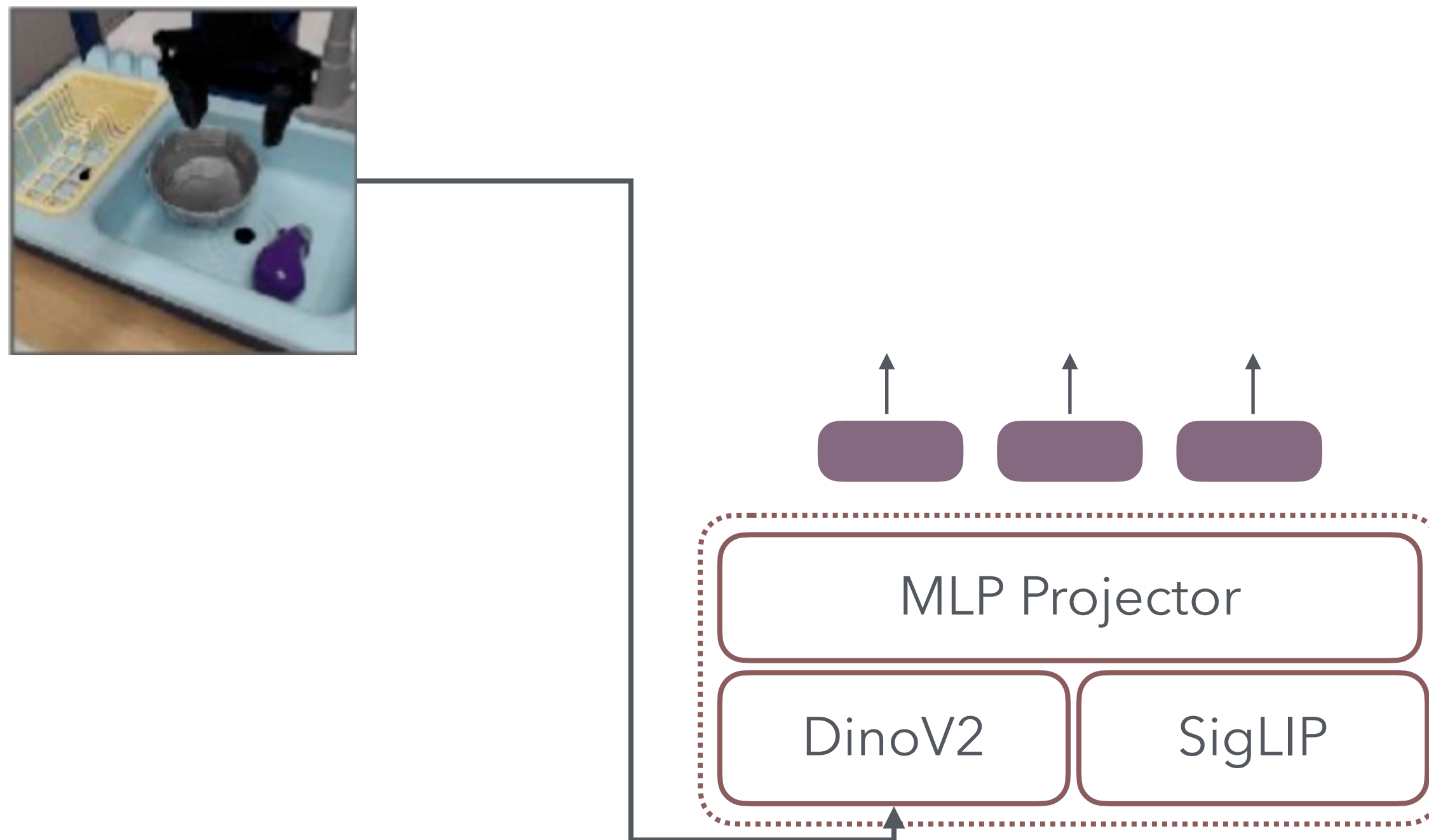
OpenVLA: Open-source Vision-Language-Action model

Deux encodeurs visuels: DinoV2 et SigLIP



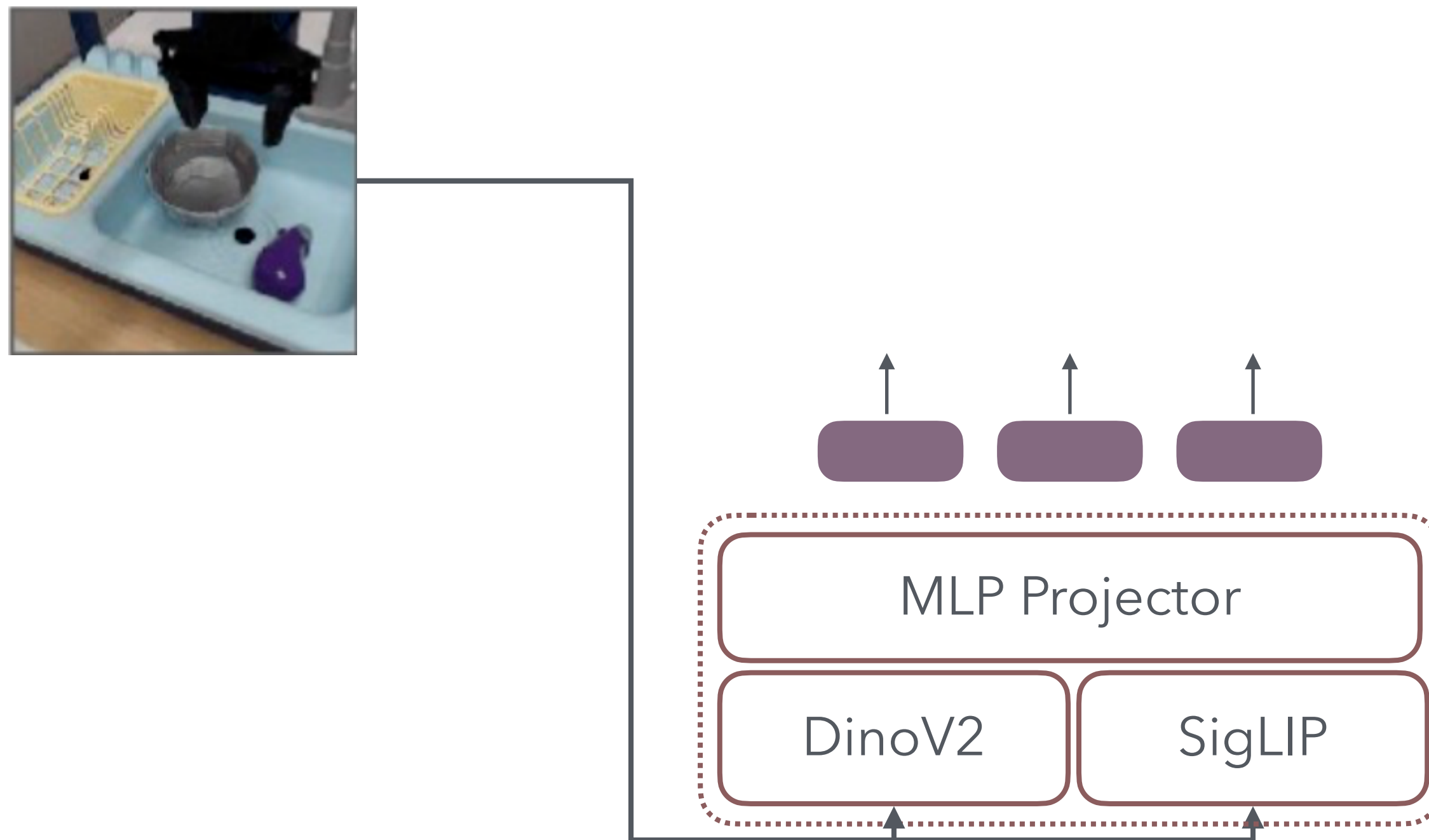
OpenVLA: Open-source Vision-Language-Action model

Deux encodeurs visuels: DinoV2 et SigLIP



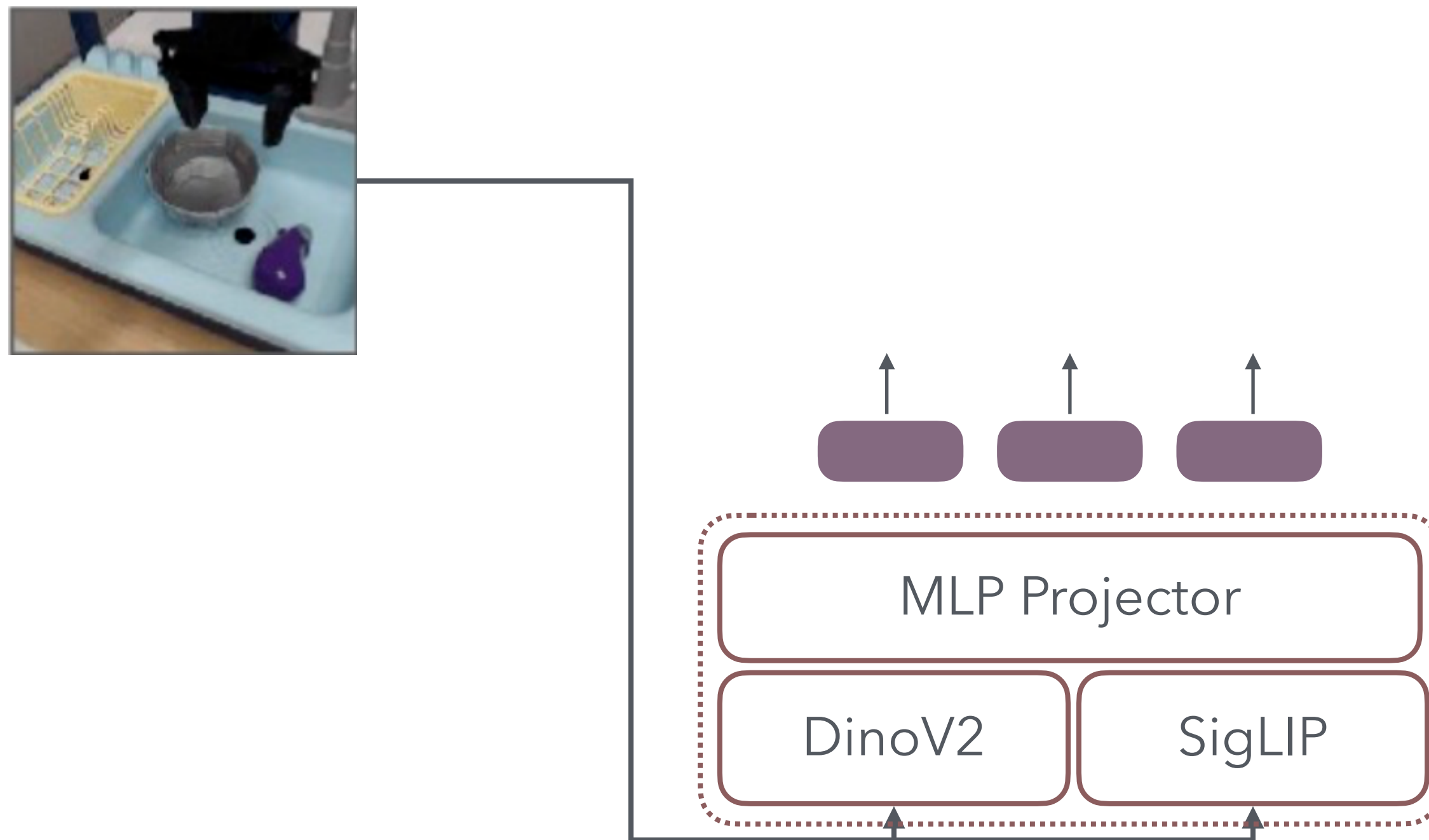
OpenVLA: Open-source Vision-Language-Action model

Deux encodeurs visuels: DinoV2 et SigLIP



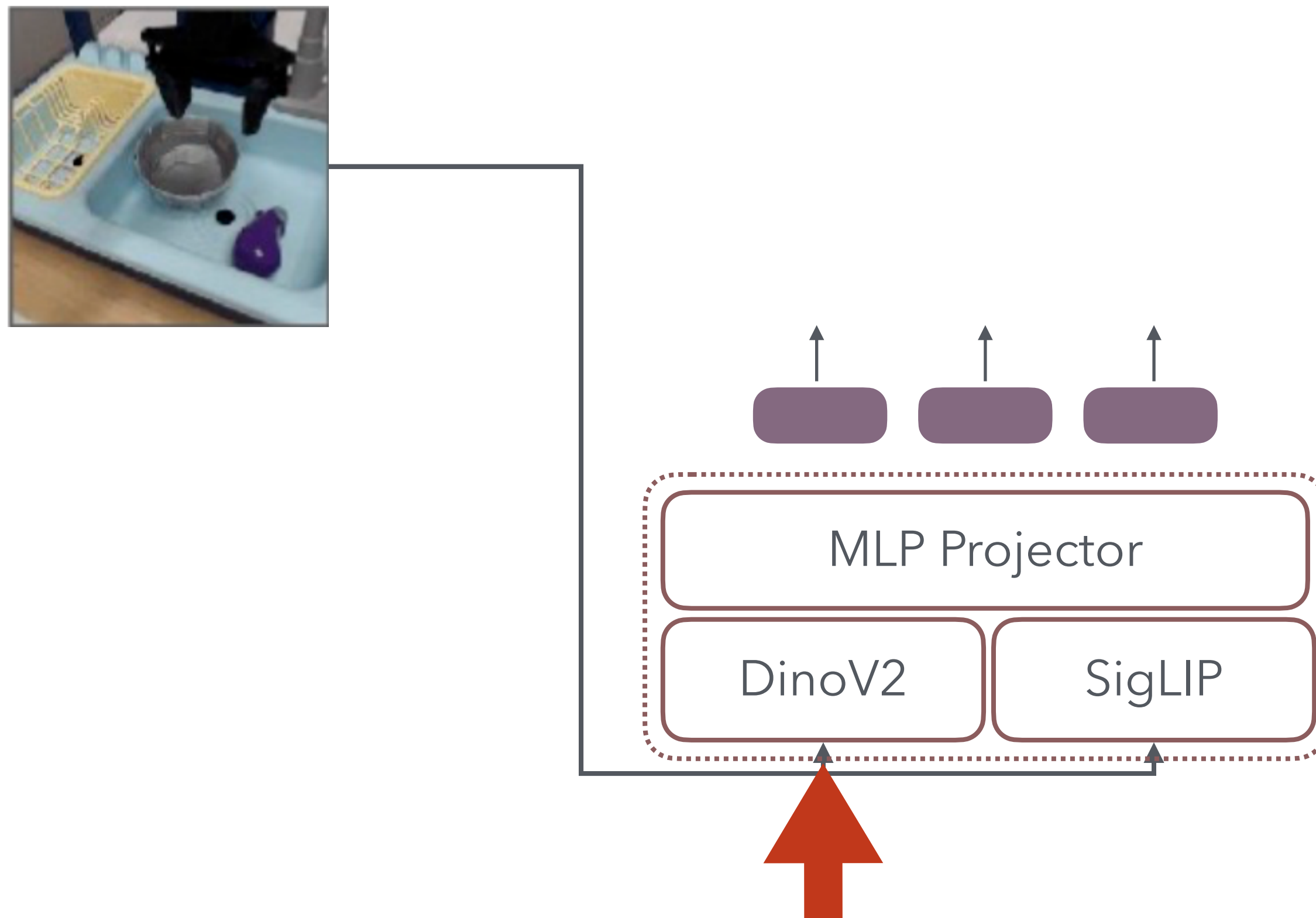
OpenVLA: Open-source Vision-Language-Action model

Deux encodeurs visuels: DinoV2 et SigLIP



OpenVLA: Open-source Vision-Language-Action model

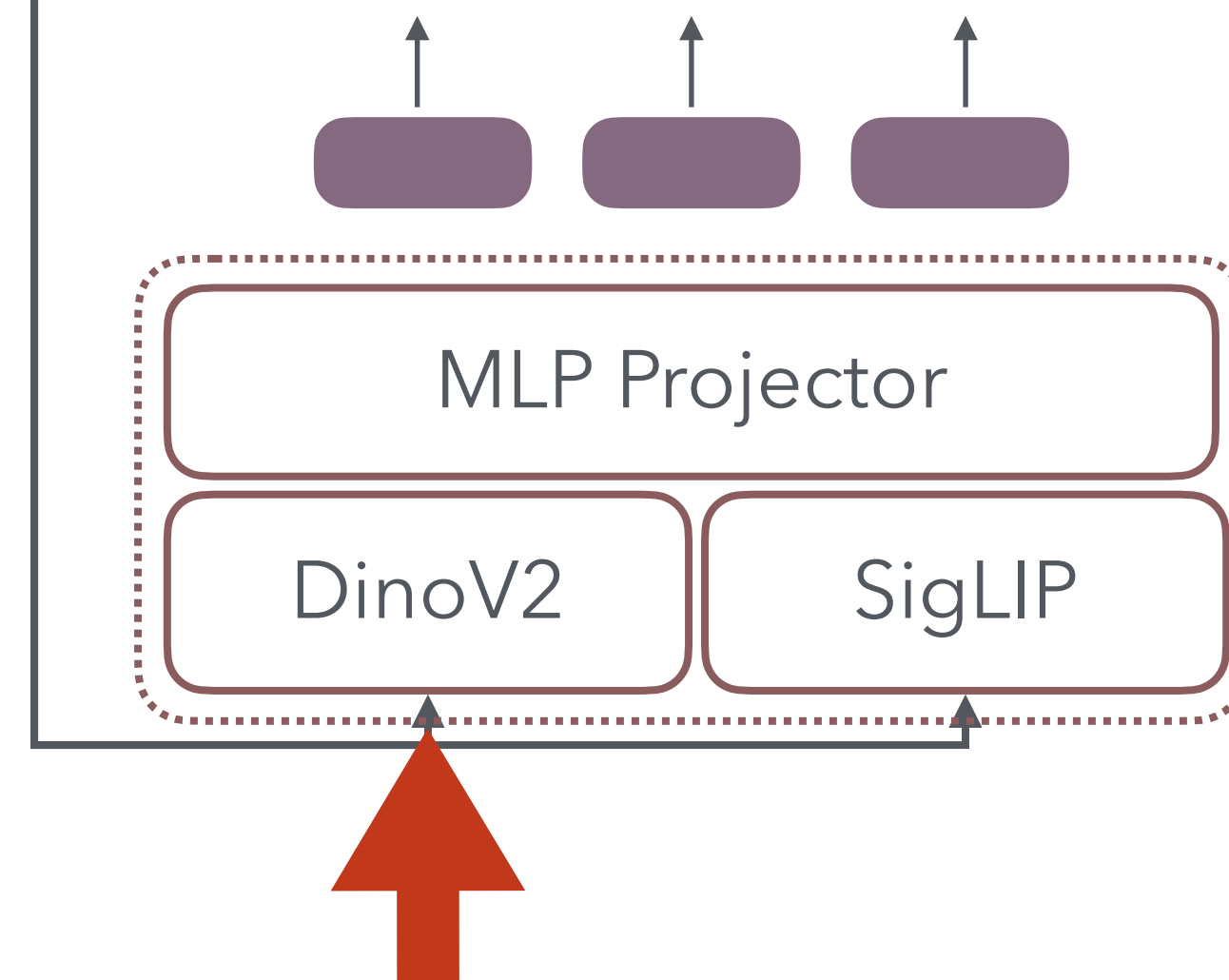
Deux encodeurs visuels: DinoV2 et SigLIP



DinoV2 ajoute des caractéristiques visuelles complémentaires utiles au raisonnement spatial.

OpenVLA: Open-source Vision-Language-Action model

Deux encodeurs visuels: DinoV2 et SigLIP

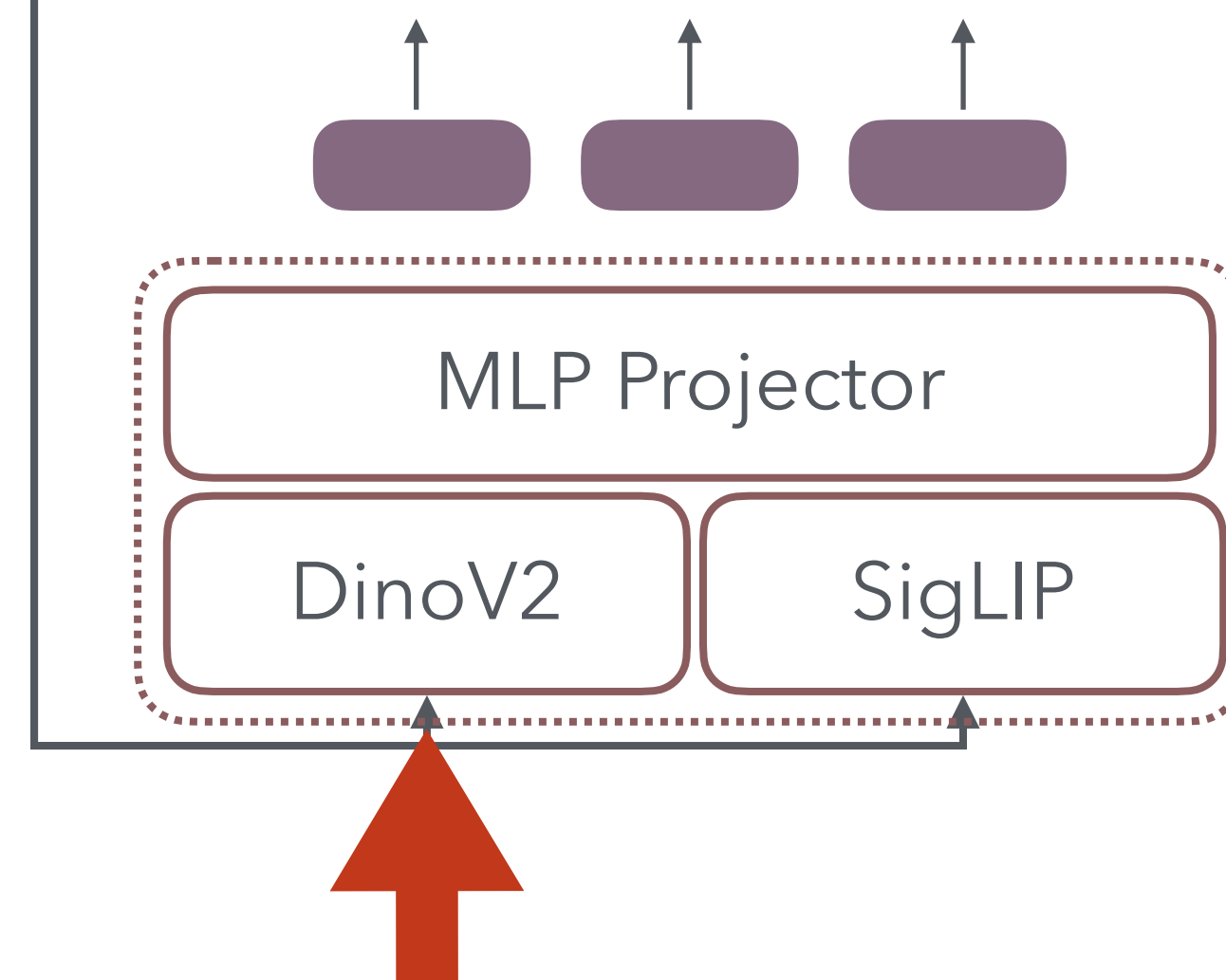


SigLIP contribue à la représentation image-langage.

DinoV2 ajoute des caractéristiques visuelles complémentaires utiles au raisonnement spatial.

OpenVLA: Open-source Vision-Language-Action model

Deux encodeurs visuels: DinoV2 et SigLIP

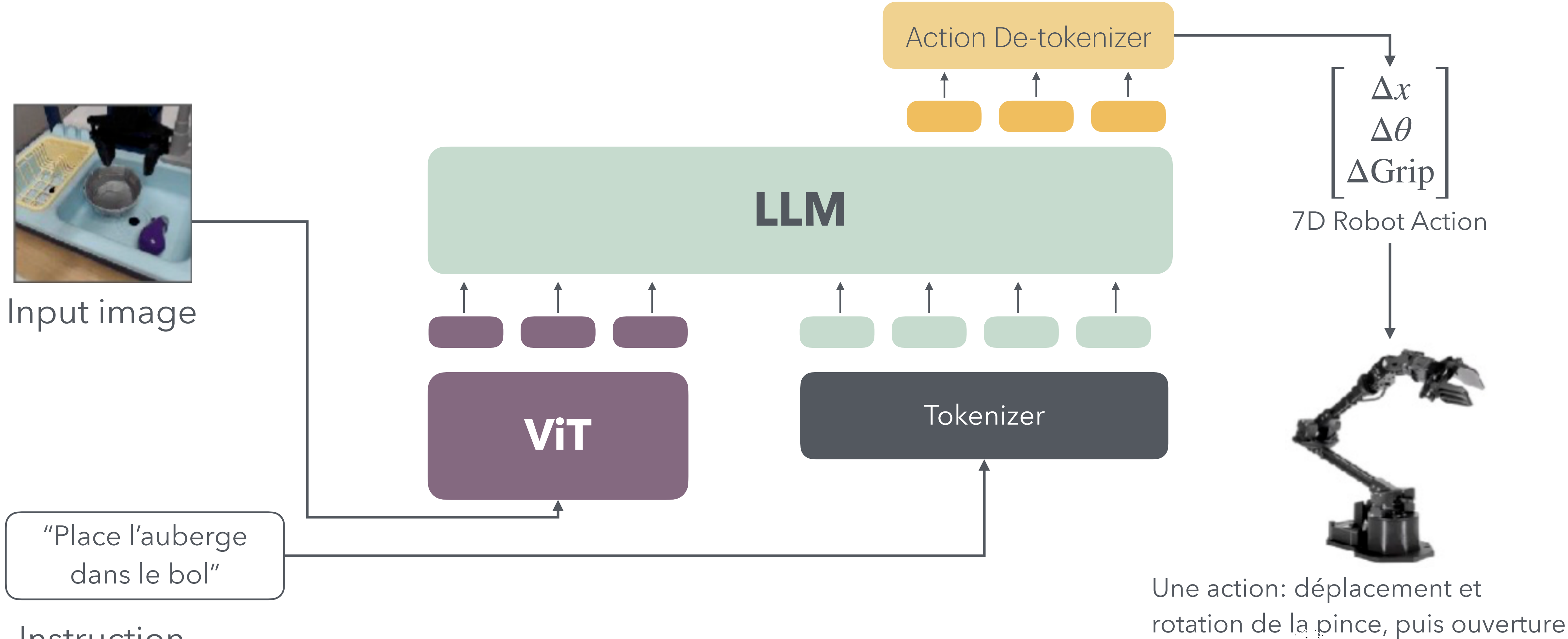


Le projecteur apprend comment ces caractéristiques visuelles doivent s'intégrer dans la représentation linguistique pré-entraînée.

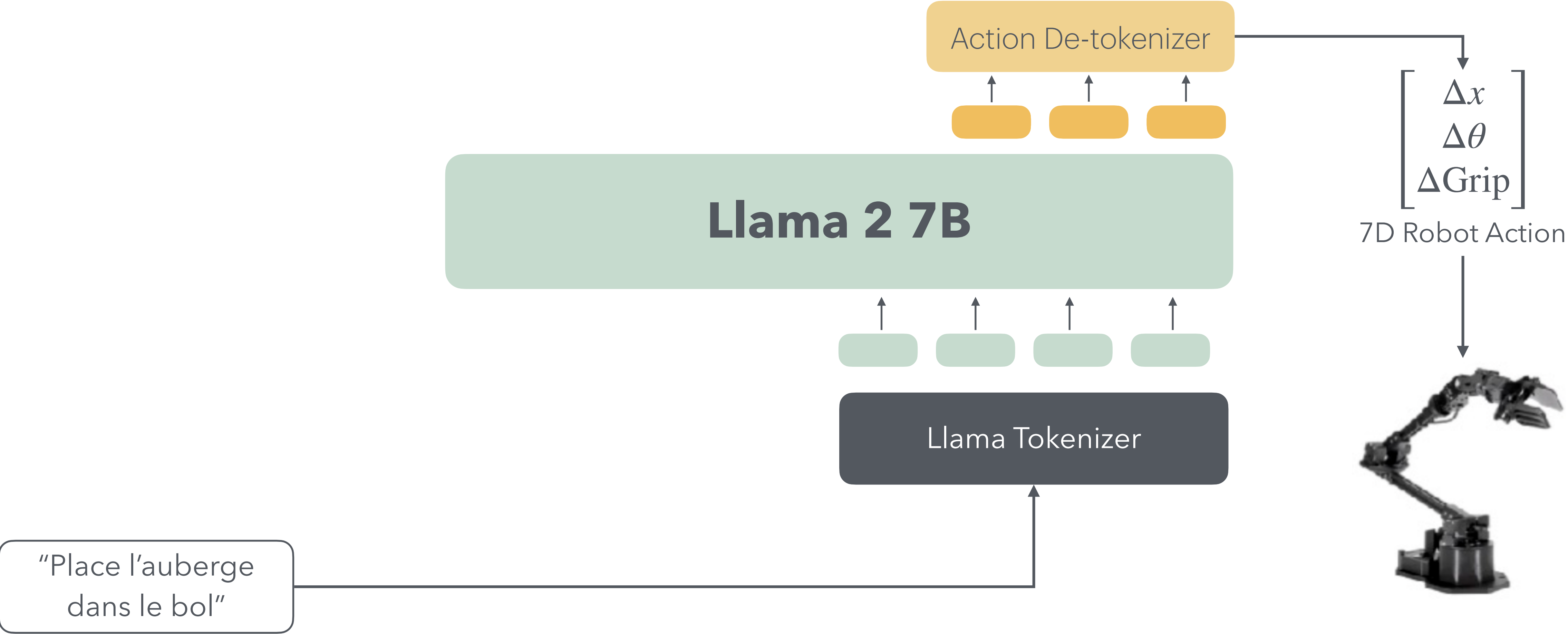
SigLIP contribue à la représentation image-langage.

DinoV2 ajoute des caractéristiques visuelles complémentaires utiles au raisonnement spatial.

OpenVLA: Open-source Vision-Language-Action model

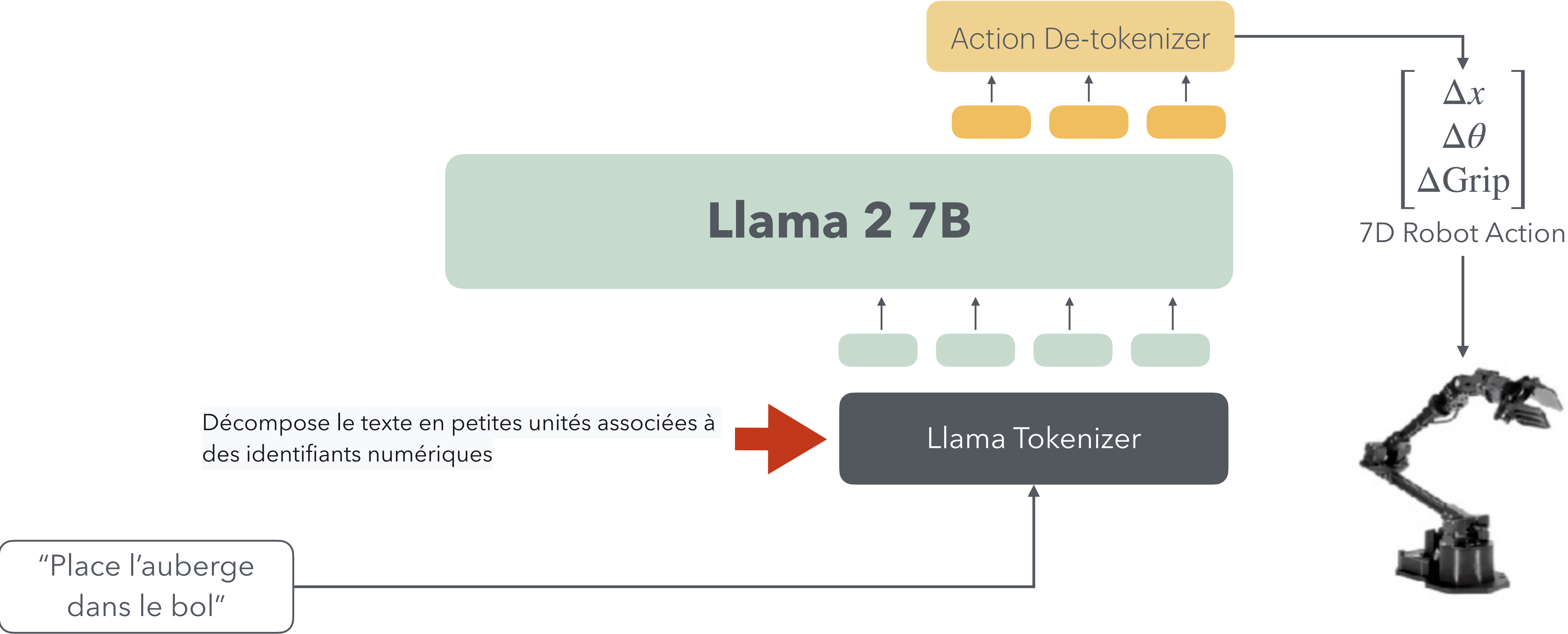


OpenVLA: Open-source Vision-Language-Action model



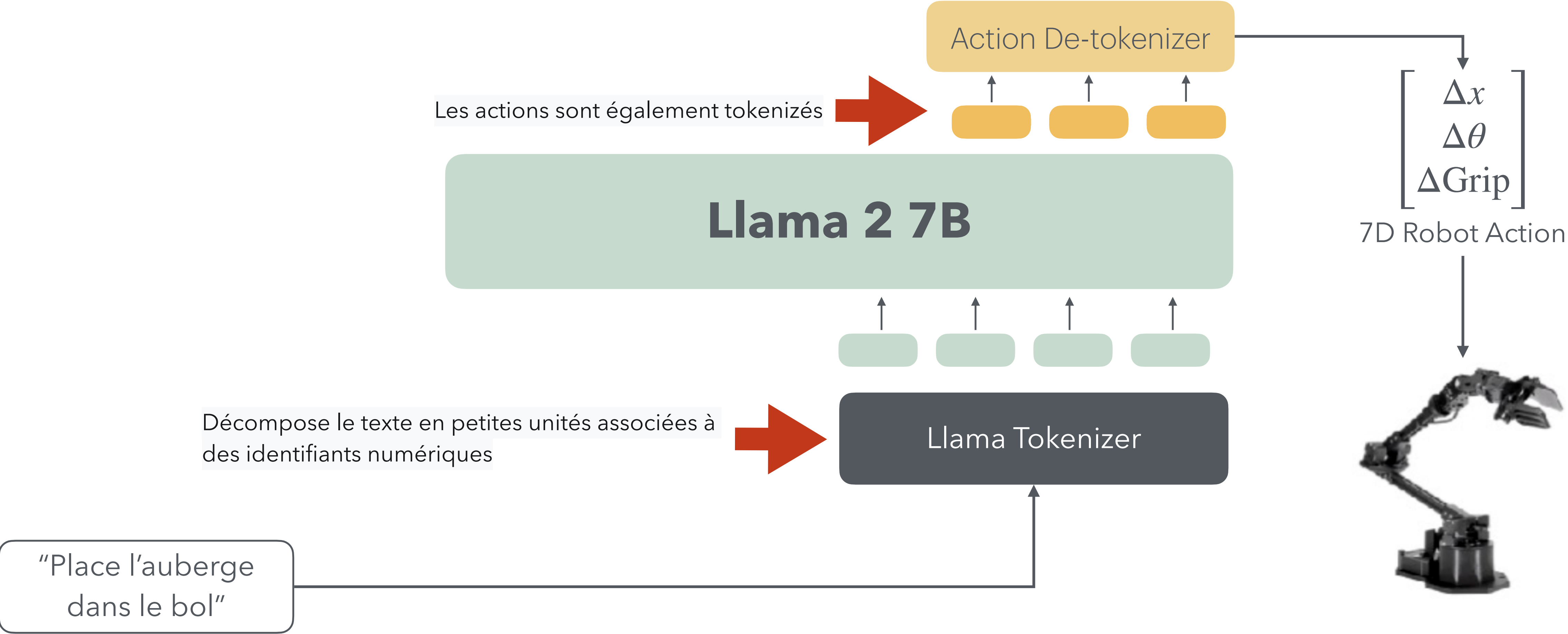
Instruction

OpenVLA: Open-source Vision-Language-Action model



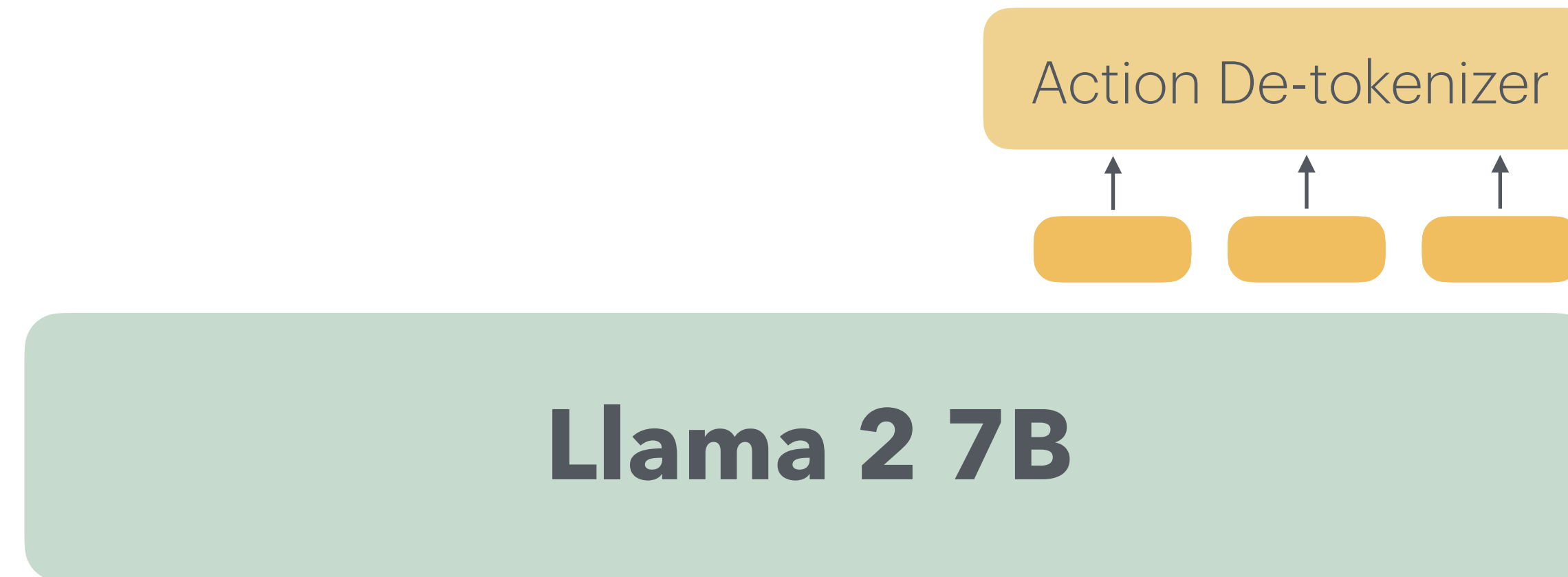
Instruction

OpenVLA: Open-source Vision-Language-Action model

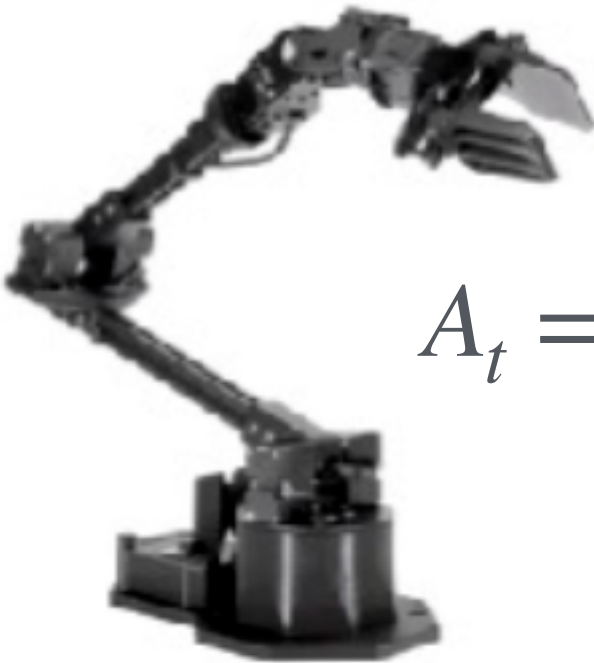
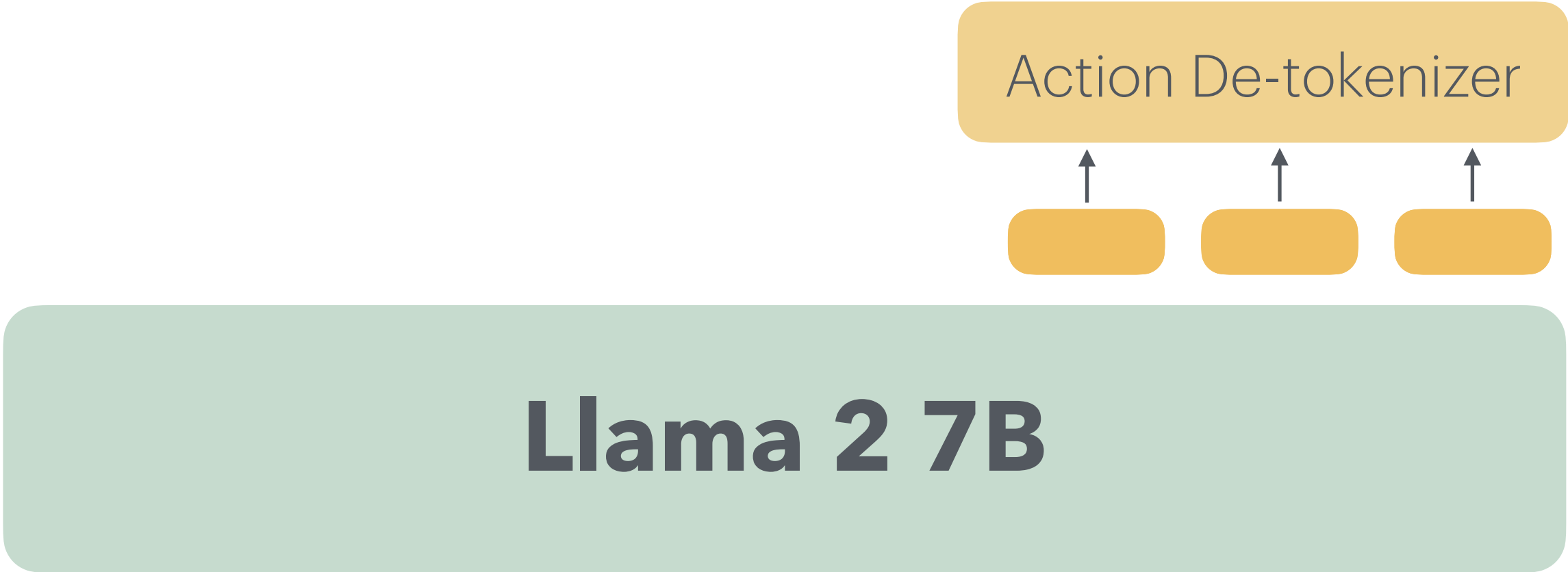


Instruction

OpenVLA: Open-source Vision-Language-Action model

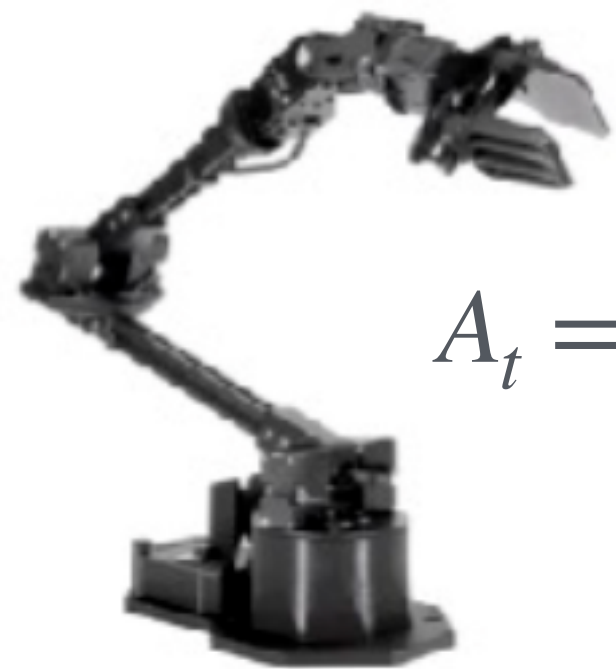
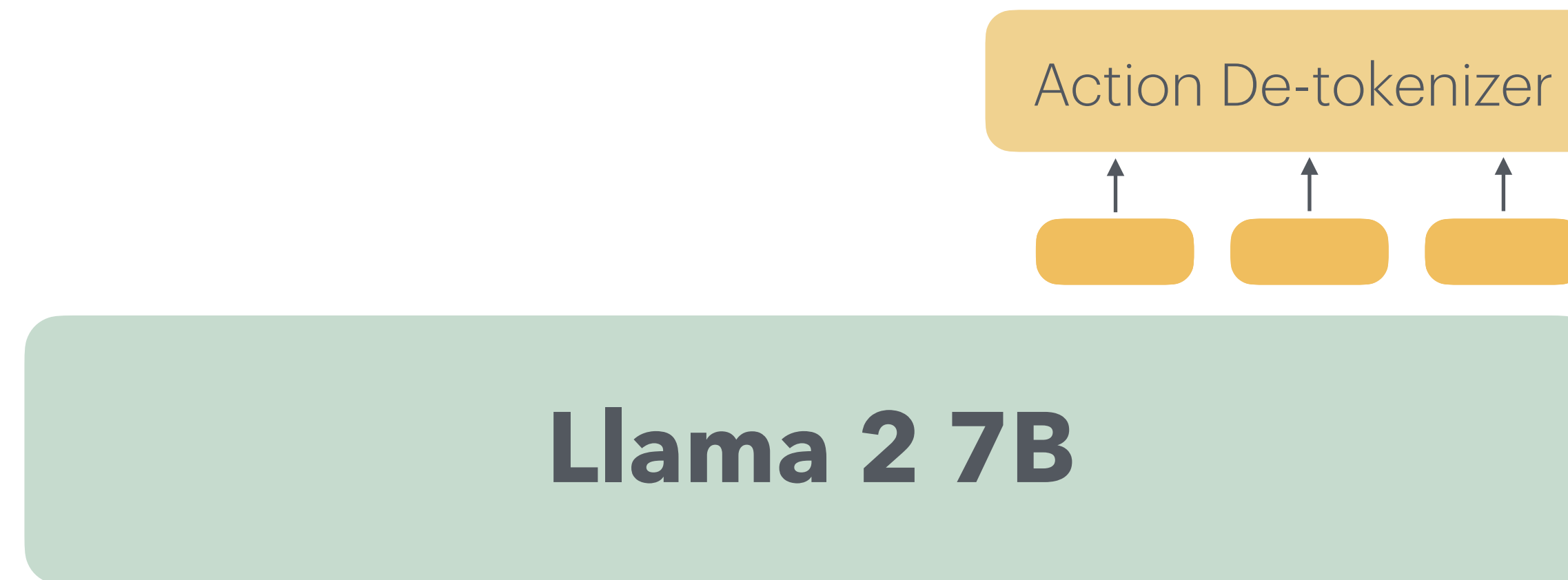


OpenVLA: Open-source Vision-Language-Action model



$$A_t = [\Delta x, \Delta y, \Delta z, \Delta r_x, \Delta r_y, \Delta r_z, g]$$

OpenVLA: Open-source Vision-Language-Action model

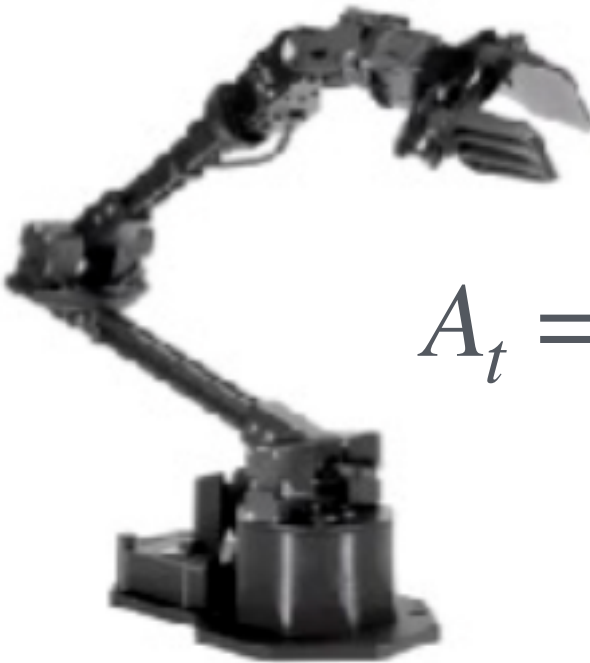
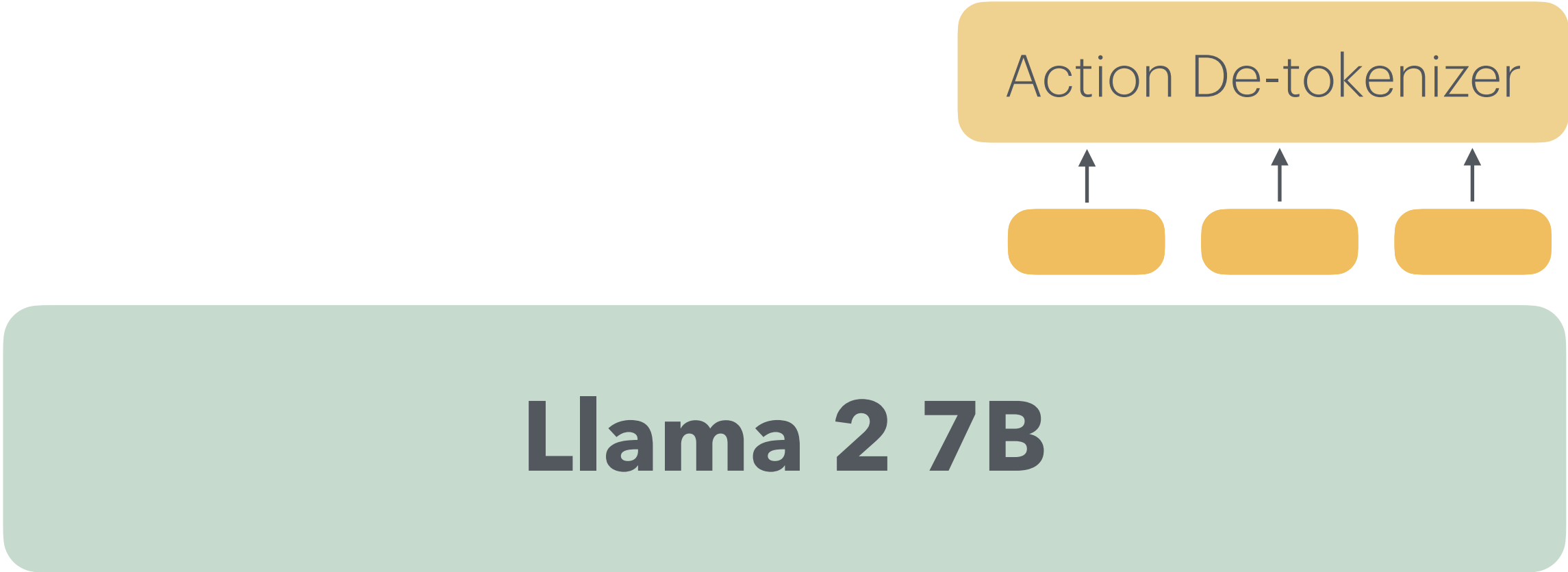


$$A_t = [\Delta x, \Delta y, \Delta z, \Delta r_x, \Delta r_y, \Delta r_z, g]$$

Valeurs continues

Ex: $\Delta x = -2.3 \text{ cm} \in [-3, -2] \text{ cm}$

OpenVLA: Open-source Vision-Language-Action model

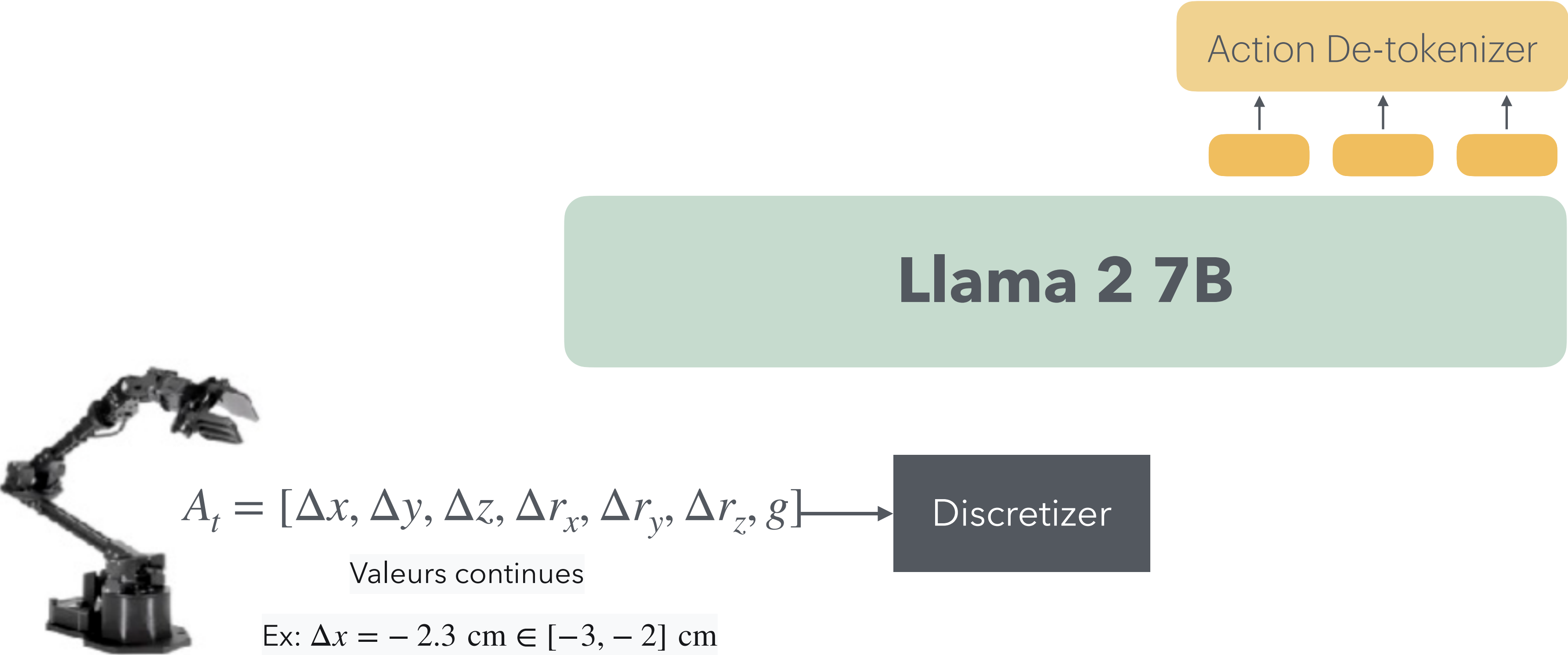


$$A_t = [\Delta x, \Delta y, \Delta z, \Delta r_x, \Delta r_y, \Delta r_z, g] \longrightarrow$$

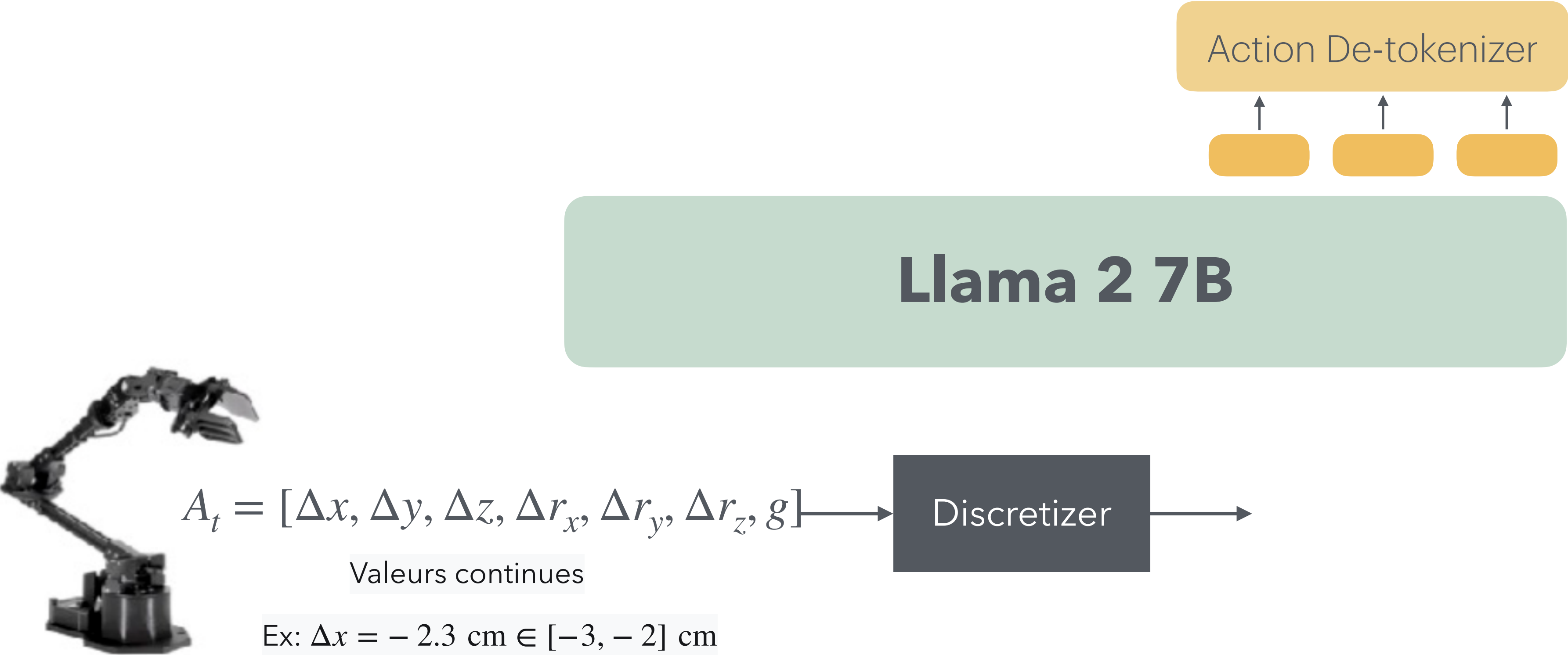
Valeurs continues

Ex: $\Delta x = -2.3 \text{ cm} \in [-3, -2] \text{ cm}$

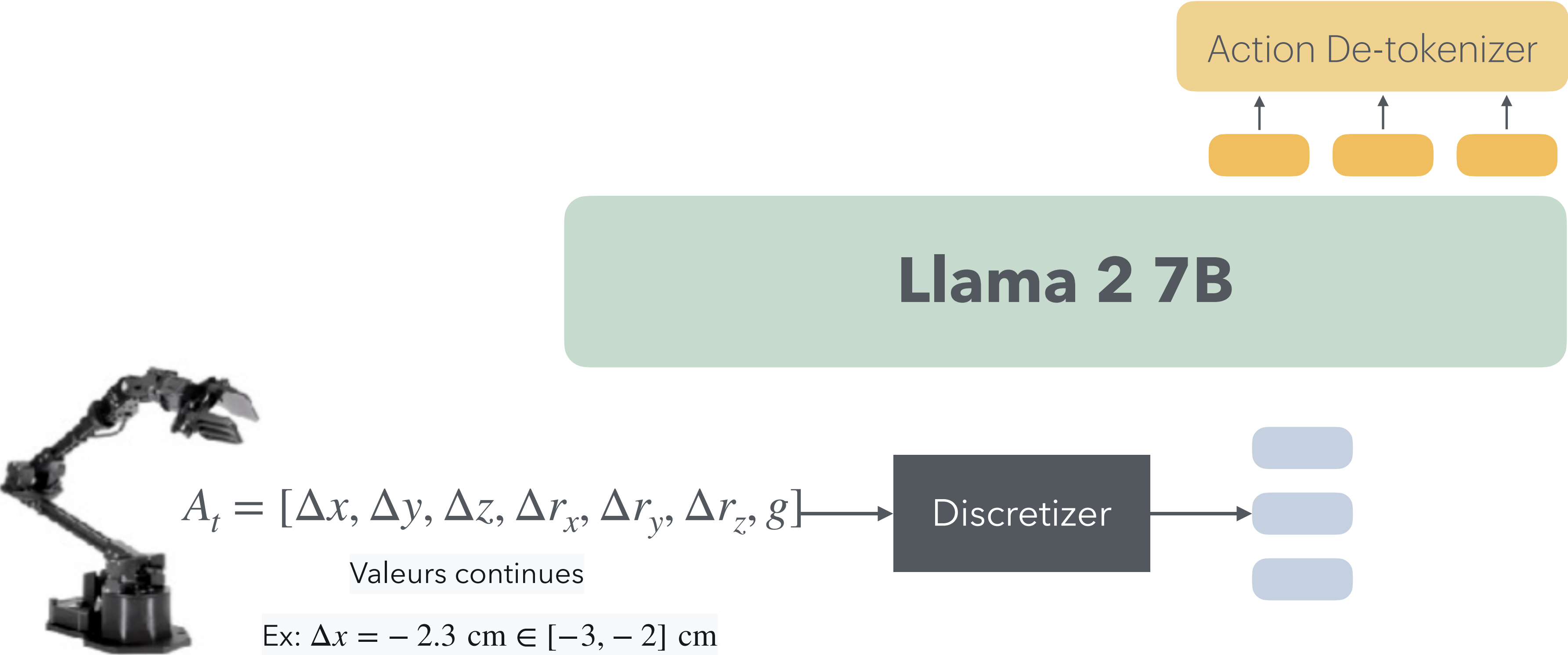
OpenVLA: Open-source Vision-Language-Action model



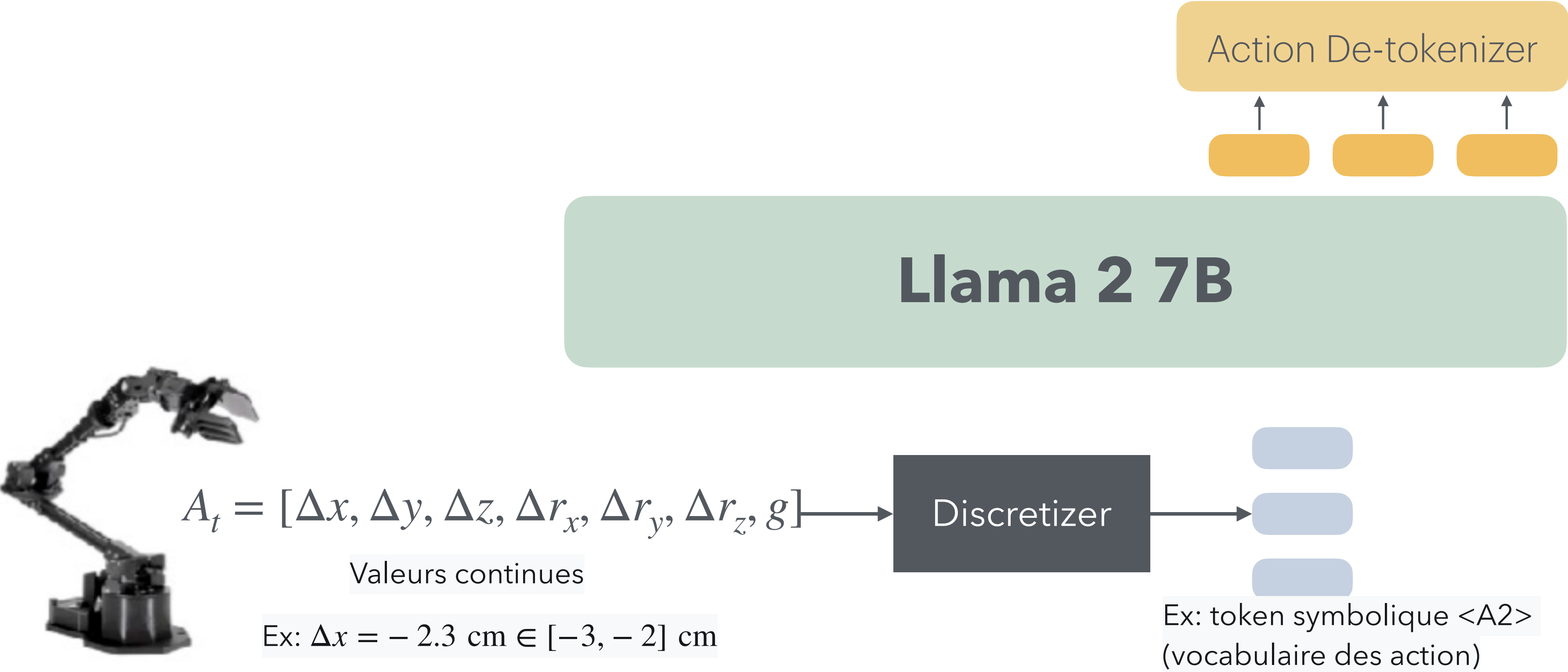
OpenVLA: Open-source Vision-Language-Action model



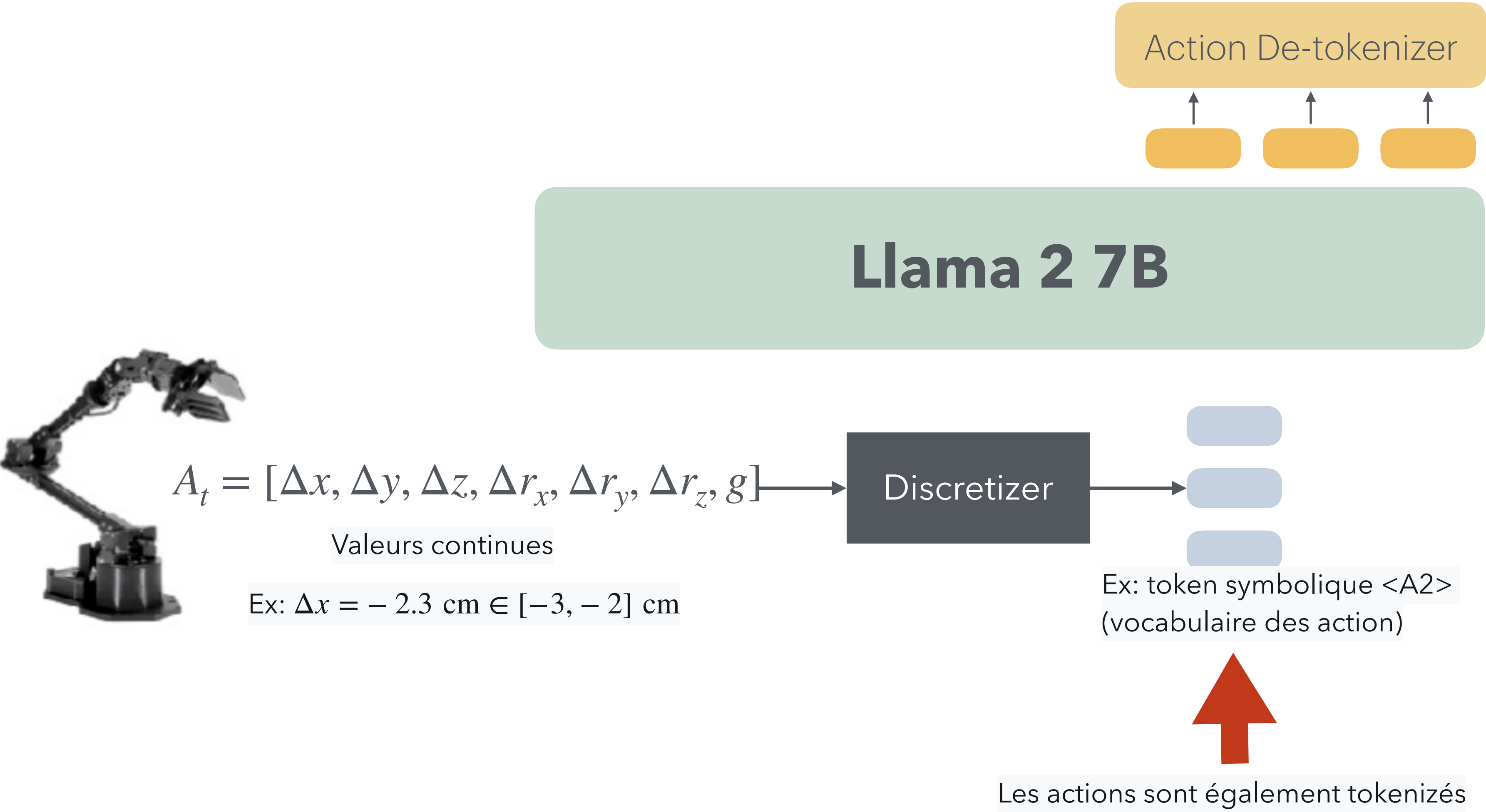
OpenVLA: Open-source Vision-Language-Action model



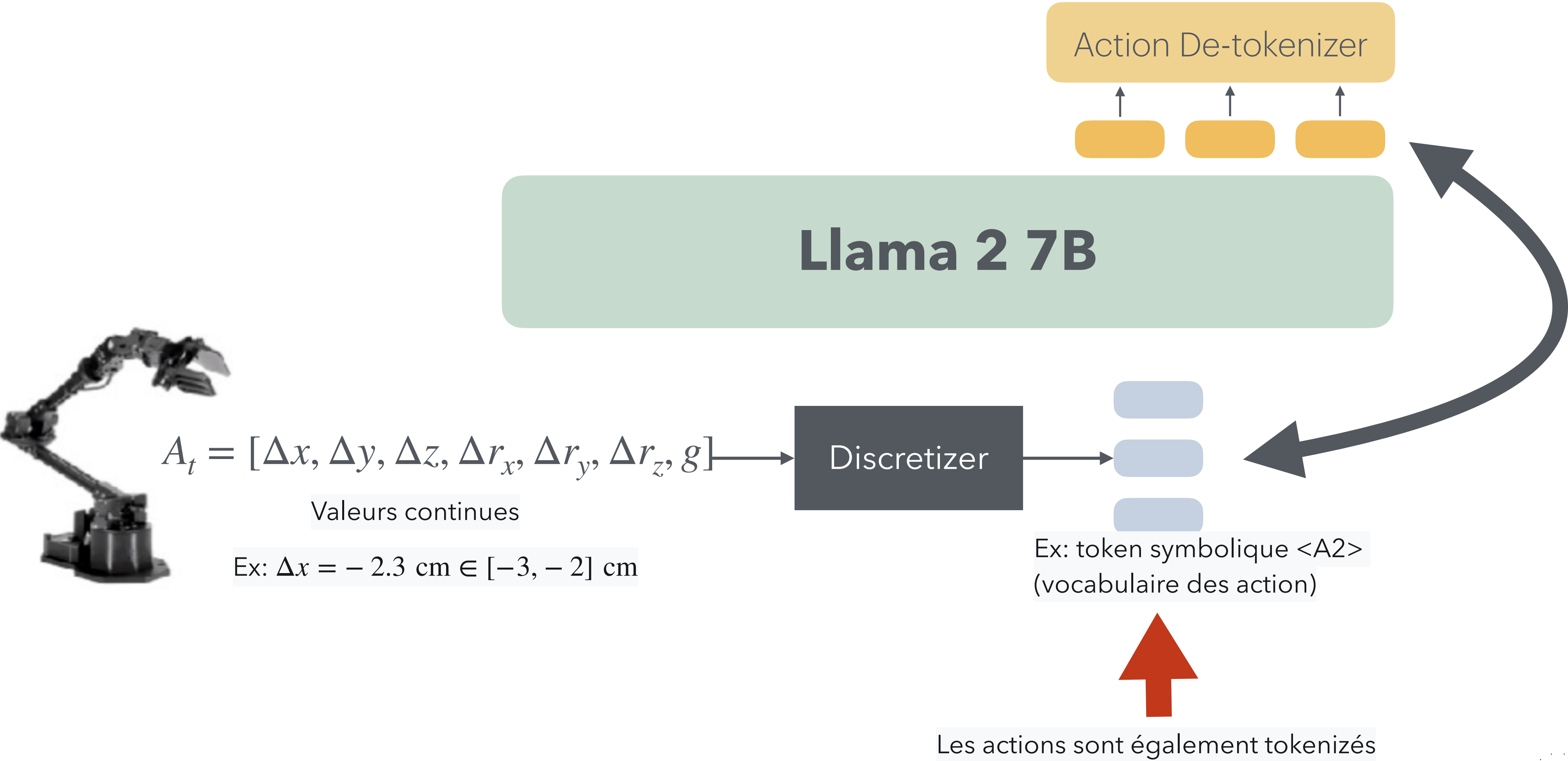
OpenVLA: Open-source Vision-Language-Action model



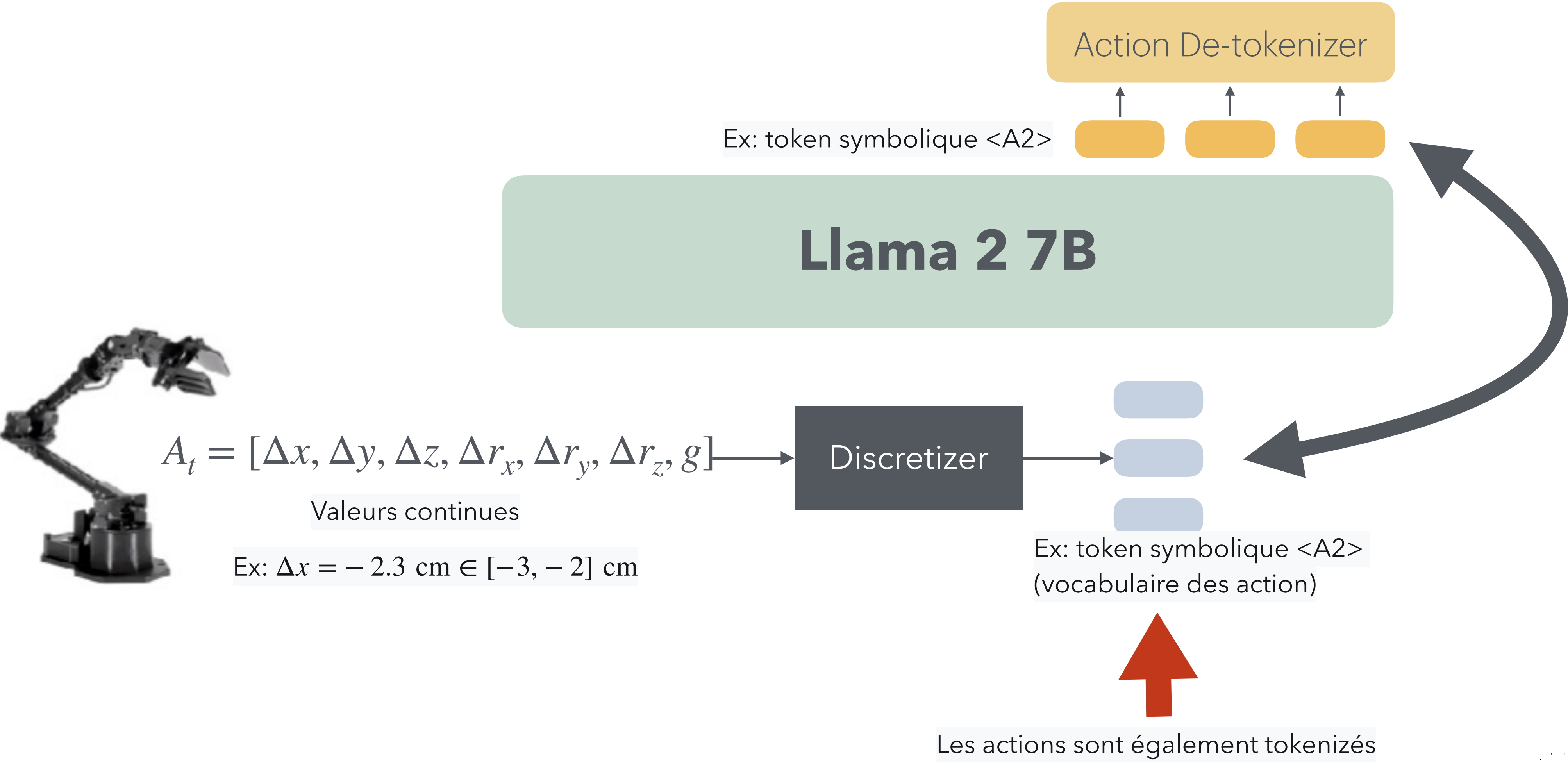
OpenVLA: Open-source Vision-Language-Action model



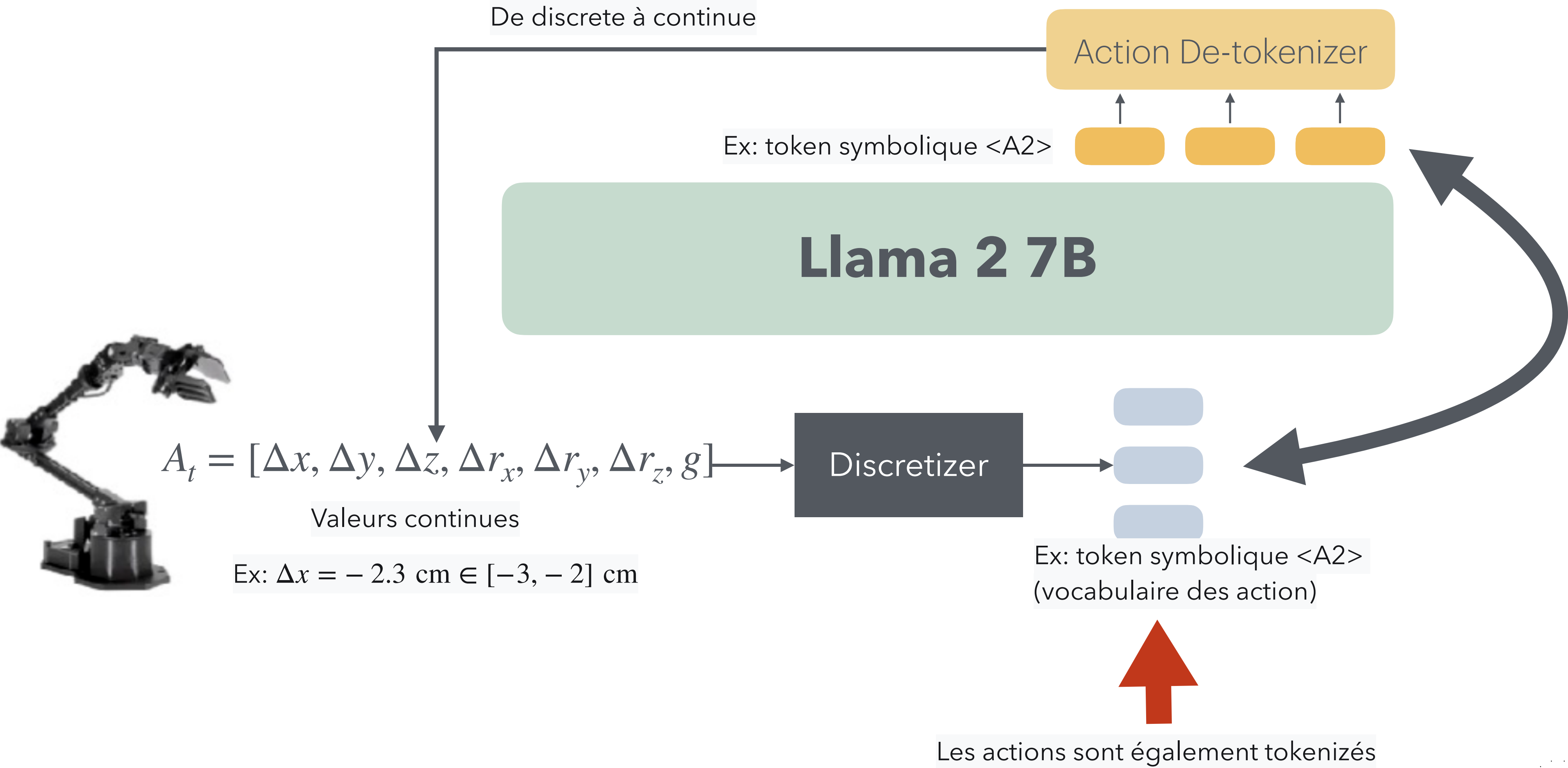
OpenVLA: Open-source Vision-Language-Action model



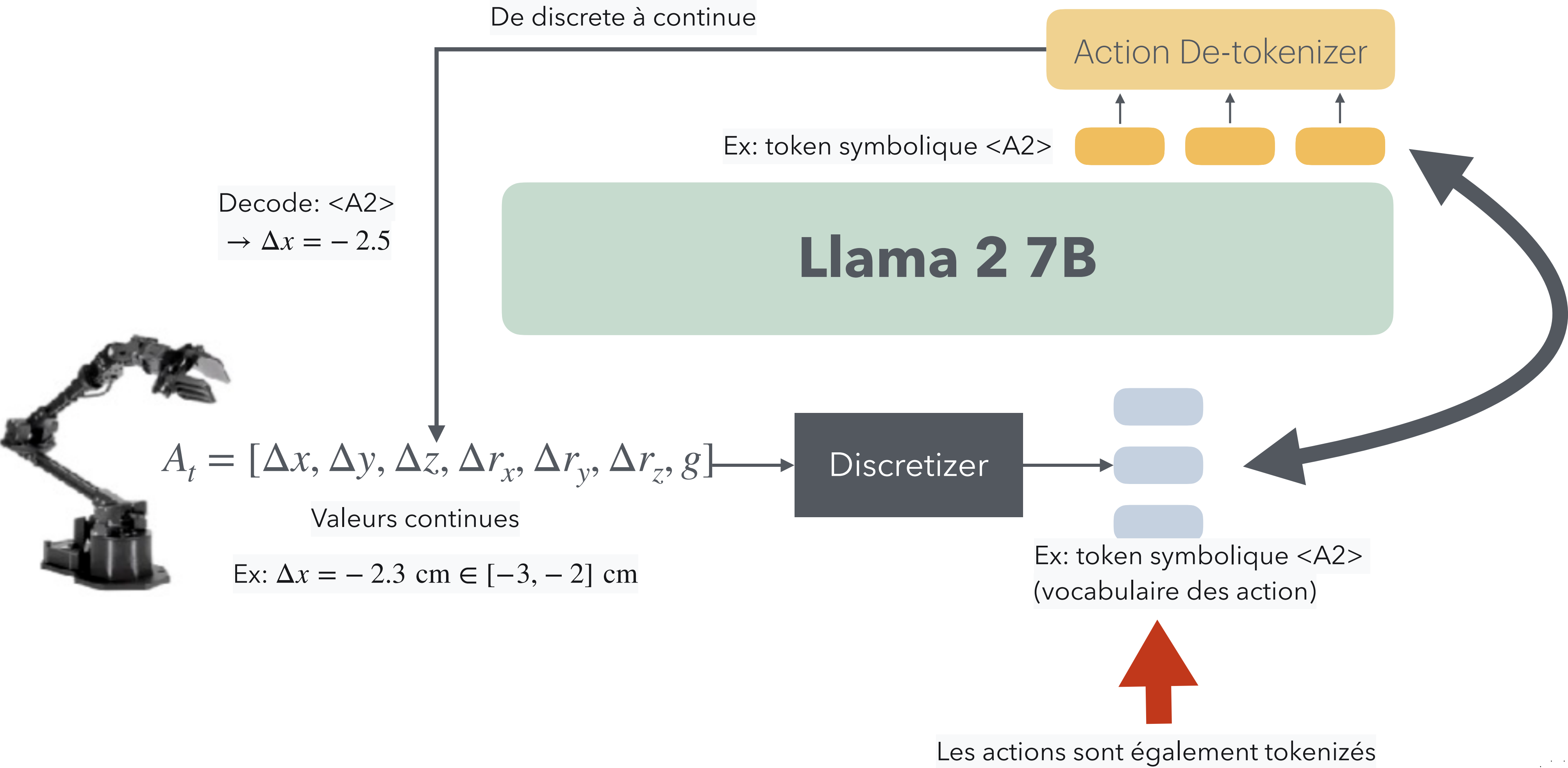
OpenVLA: Open-source Vision-Language-Action model



OpenVLA: Open-source Vision-Language-Action model



OpenVLA: Open-source Vision-Language-Action model



OpenVLA: Open-source Vision-Language-Action model

Comment ce genre de modèle est-il entraîné ou pré-entraîné ?

Imitation Learning

Comment ce genre de modèle est-il entraîné ou pré-entraîné ?

Imitation Learning

OERL Framework

The framework consists of the following components and flow:

- Task Learning** and **Data Collection** feed into **Task Generalization** and **Data Distribution**.
- Task Generalization** leads to **Task Distribution**.
- Data Distribution** leads to **Data Collection**.

22 Embodiments

The framework supports 22 different robot embodiments, including:

- QT-Opt, TOTO, RT-1, and others.

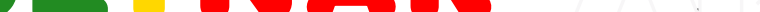
60 Datasets

The framework supports 60 different datasets, including:

- pick something, pour, stack, route, pick green chip bag, and others.

Figure 1 consists of five subplots (a-e) illustrating the dataset overview:

- (a) **# Datasets per Robot Embodiment**: A bar chart showing the number of datasets for various robot embodiments. Franka is the most common with 7 datasets, followed by Sawyer (6), and Kinova Gen3 (5). Other robots like Google Robot, vArm, and WidowX have 4 datasets each.
- (b) **# Scenes per Embodiment**: A pie chart showing the distribution of scenes across robot embodiments. Franka has the largest share (blue), followed by Google Robot (orange) and vArm (green).
- (c) **# Trajectories per Embodiment**: A pie chart showing the distribution of trajectories. vArm (green) and Google Robot (orange) have the highest counts, followed by Franka (blue) and Kuka iwa (pink).
- (d) **Common Dataset Skills**: A bar chart showing the frequency of various skills across the dataset. Skills like 'Pick' and 'Place' are the most common, with 'Pick' being the most frequent skill.
- (e) **Common Dataset Objects**: A bar chart showing the frequency of various objects. Objects like 'Cup', 'Bottle', and 'Box' are common, with 'Cup' being the most frequent object.



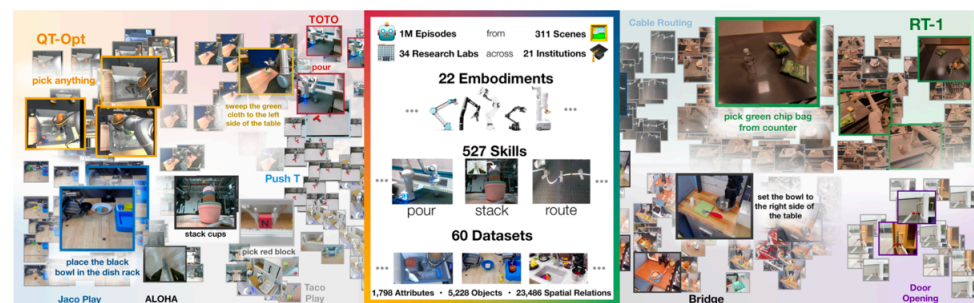
OpenVLA: Open-source Vision-Language-Action model

Comment ce genre de modèle est-il entraîné ou pré-entraîné ?

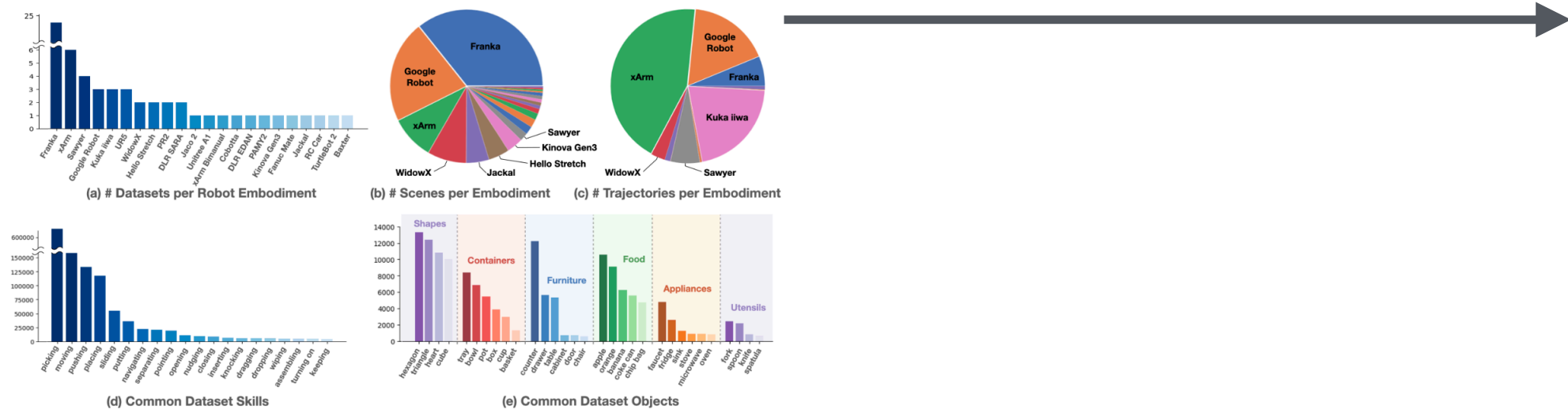
Collection d’images, text, action
robotiques

Imitation Learning

Dataset Overview



We introduce the Open X-Embodiment Dataset, the largest open-source real robot dataset to date. It contains 1M+ real robot trajectories spanning 22 robot embodiments, from single robot arms to bi-manual robots and quadrupeds.



Open X-Embodiement
(<https://robotics-transformer-x.github.io/>)

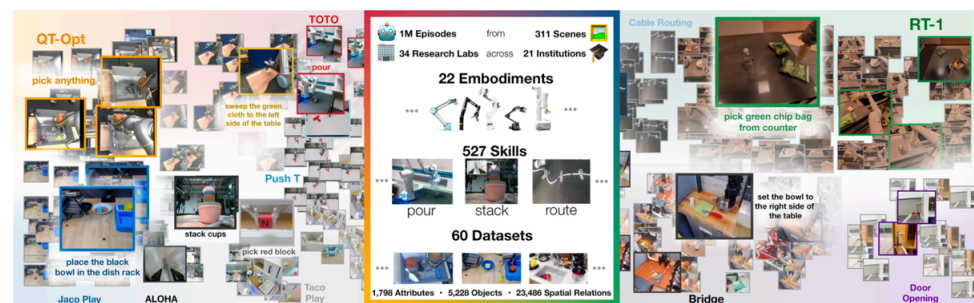
OpenVLA: Open-source Vision-Language-Action model

Comment ce genre de modèle est-il entraîné ou pré-entraîné ?

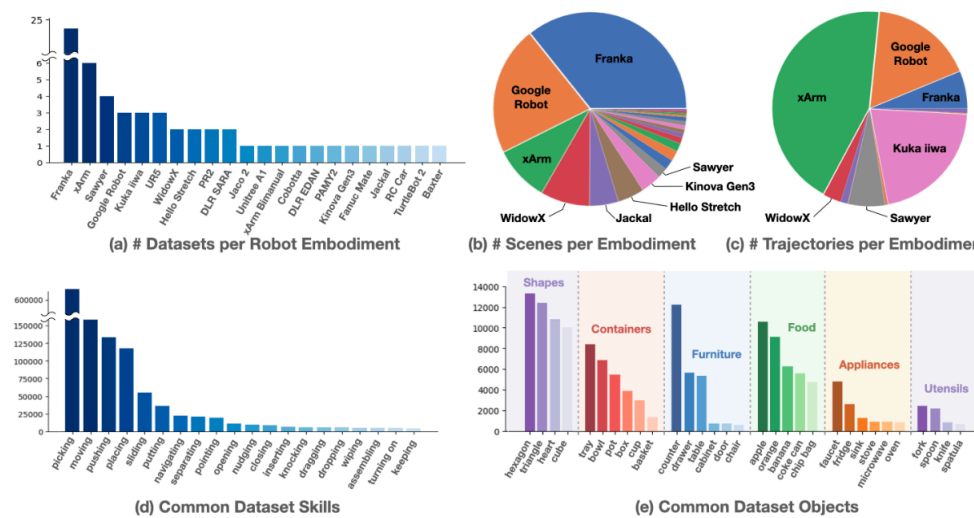
Collection d’images, text, action
robotiques

Imitation Learning

Dataset Overview



We introduce the Open X-Embodiment Dataset, the largest open-source real robot dataset to date. It contains 1M+ real robot trajectories spanning 22 robot embodiments, from single robot arms to bi-manual robots and quadrupeds.



Open X-Embodiement
(<https://robotics-transformer-x.github.io/>)

Filtrer en fonction de
la tache a apprendre

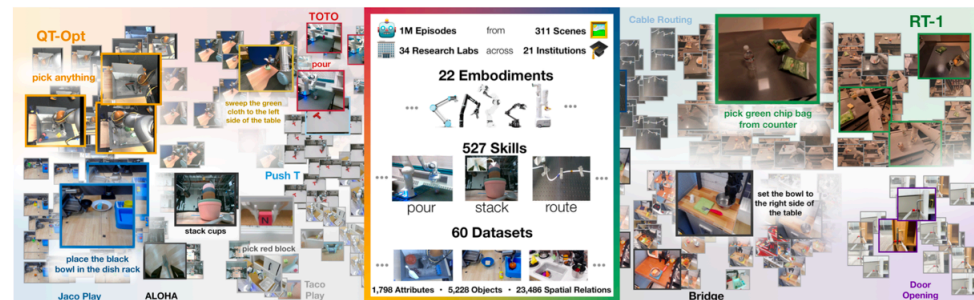
OpenVLA: Open-source Vision-Language-Action model

Comment ce genre de modèle est-il entraîné ou pré-entraîné ?

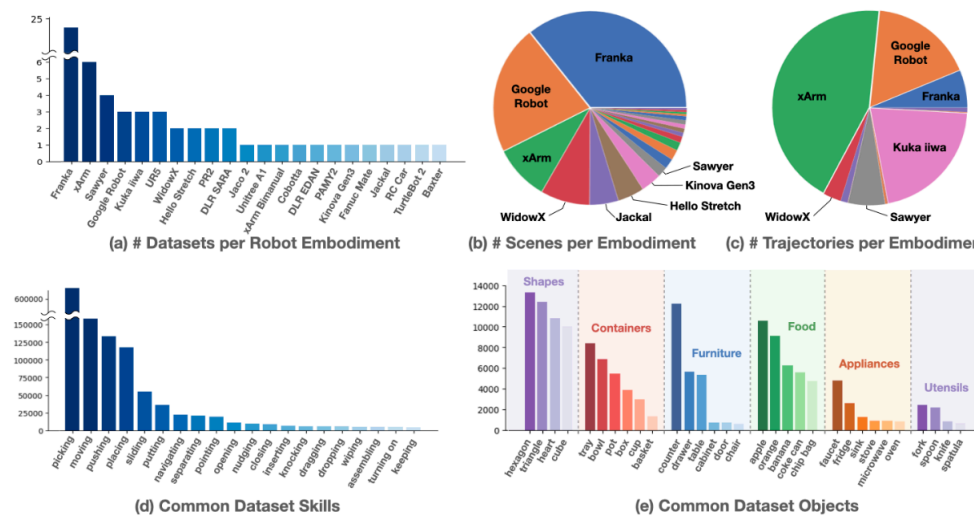
Collection d’images, text, action
robotiques

Imitation Learning

Dataset Overview



We introduce the Open X-Embodiment Dataset, the largest open-source real robot dataset to date. It contains 1M+ real robot trajectories spanning 22 robot embodiments, from single robot arms to bi-manual robots and quadrupeds.



Open X-Embodiement
(<https://robotics-transformer-x.github.io/>)

Filtrer en fonction de
la tache a apprendre

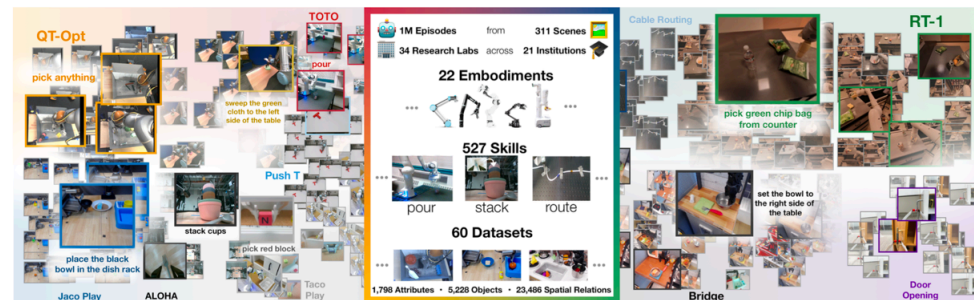
OpenVLA: Open-source Vision-Language-Action model

Comment ce genre de modèle est-il entraîné ou pré-entraîné ?

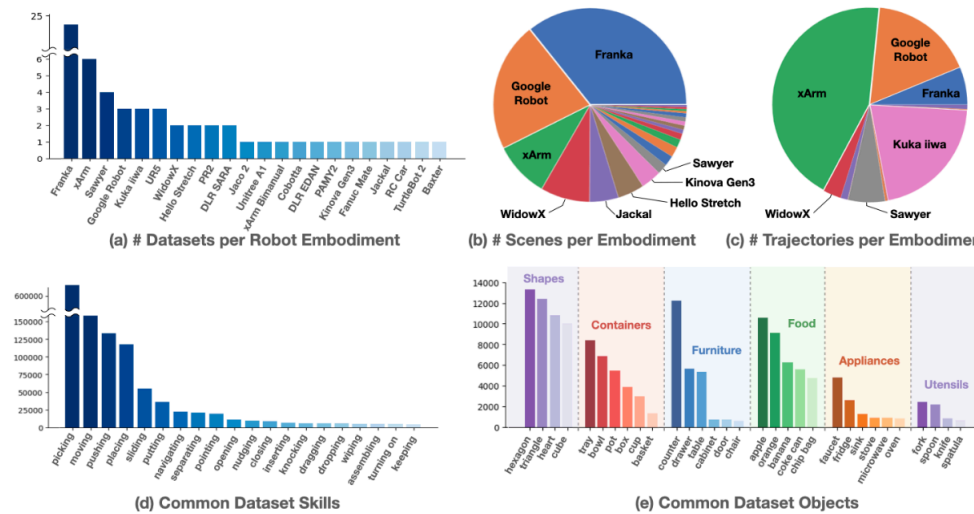
Collection d'images, text, action
robotiques

Imitation Learning

Dataset Overview



We introduce the Open X-Embodiment Dataset, the largest open-source real robot dataset to date. It contains 1M+ real robot trajectories spanning 22 robot embodiments, from single robot arms to bi-manual robots and quadrupeds.



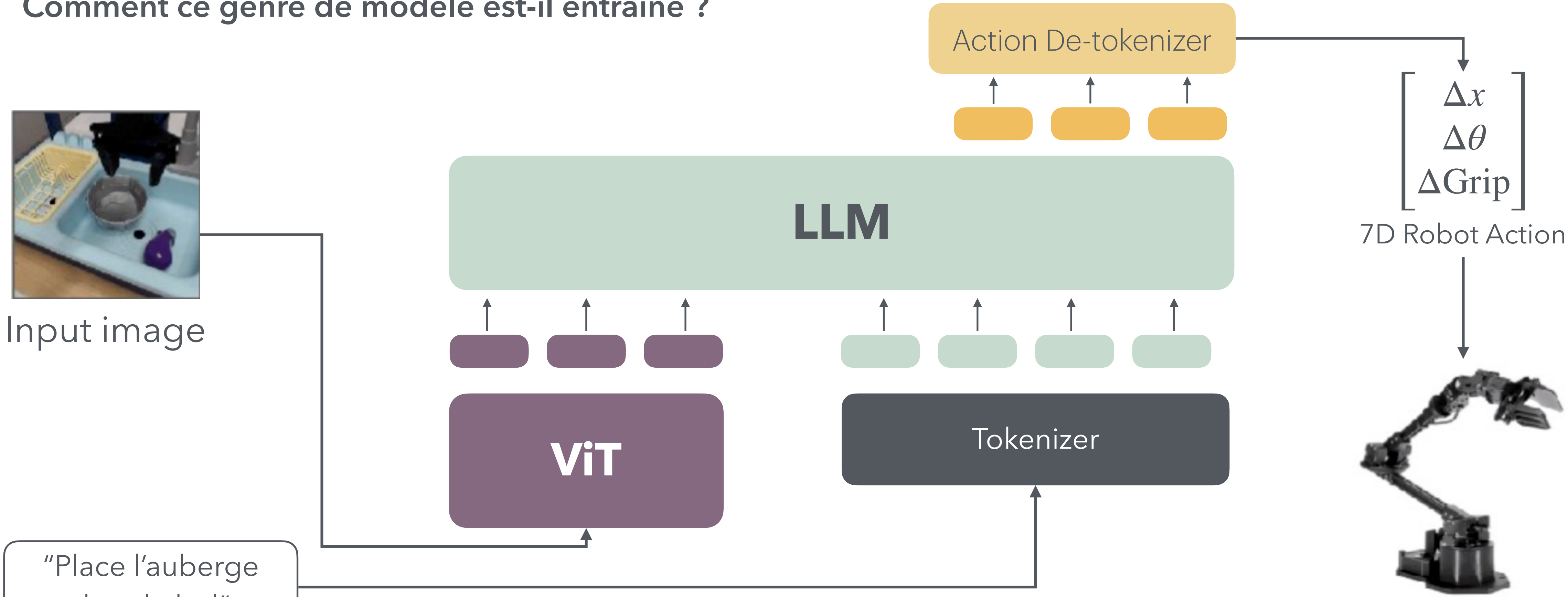
Open X-Embodiement
(<https://robotics-transformer-x.github.io/>)

Base de données
sélectionnées

Filtrer en fonction de
la tache a apprendre

OpenVLA: Open-source Vision-Language-Action model

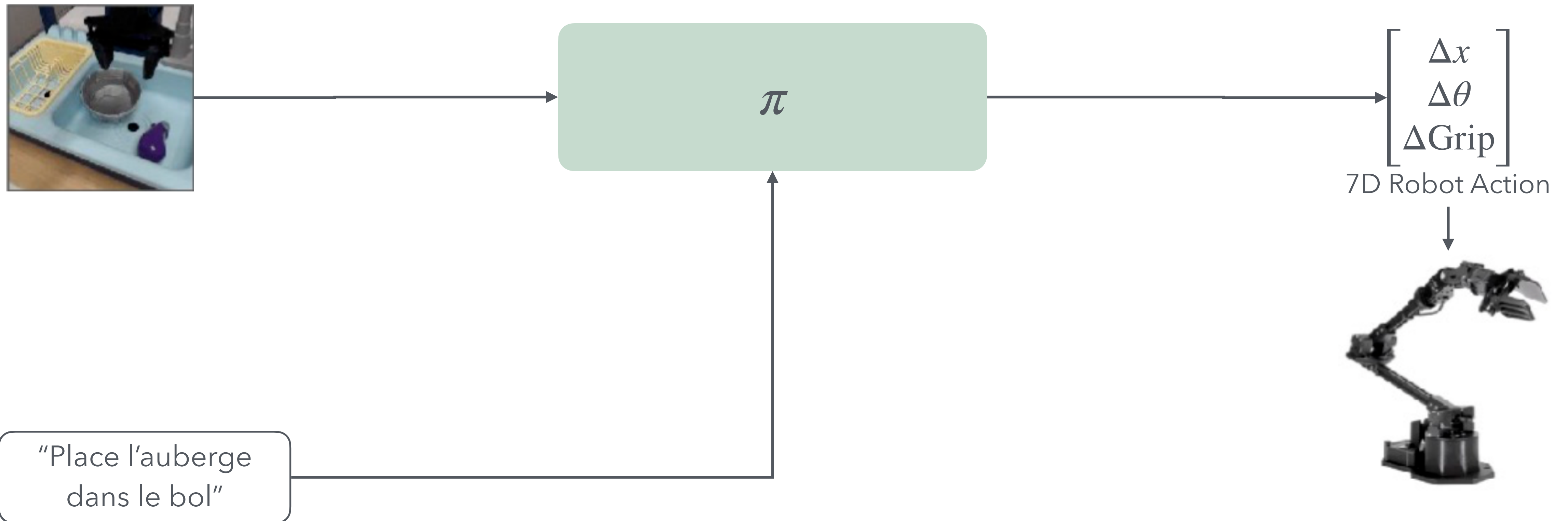
Comment ce genre de modèle est-il entraîné ?



Une action: déplacement et rotation de la pince, puis ouverture

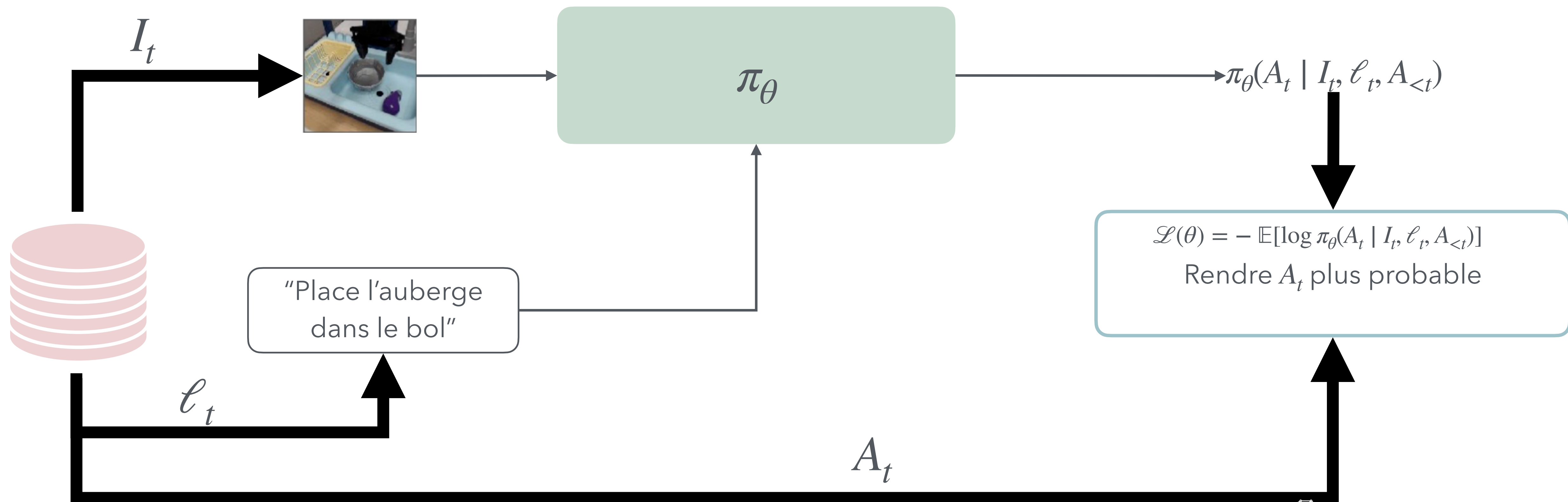
OpenVLA: Open-source Vision-Language-Action model

Comment ce genre de modèle est-il entraîné ou pré-entraîné ?



OpenVLA: Open-source Vision-Language-Action model

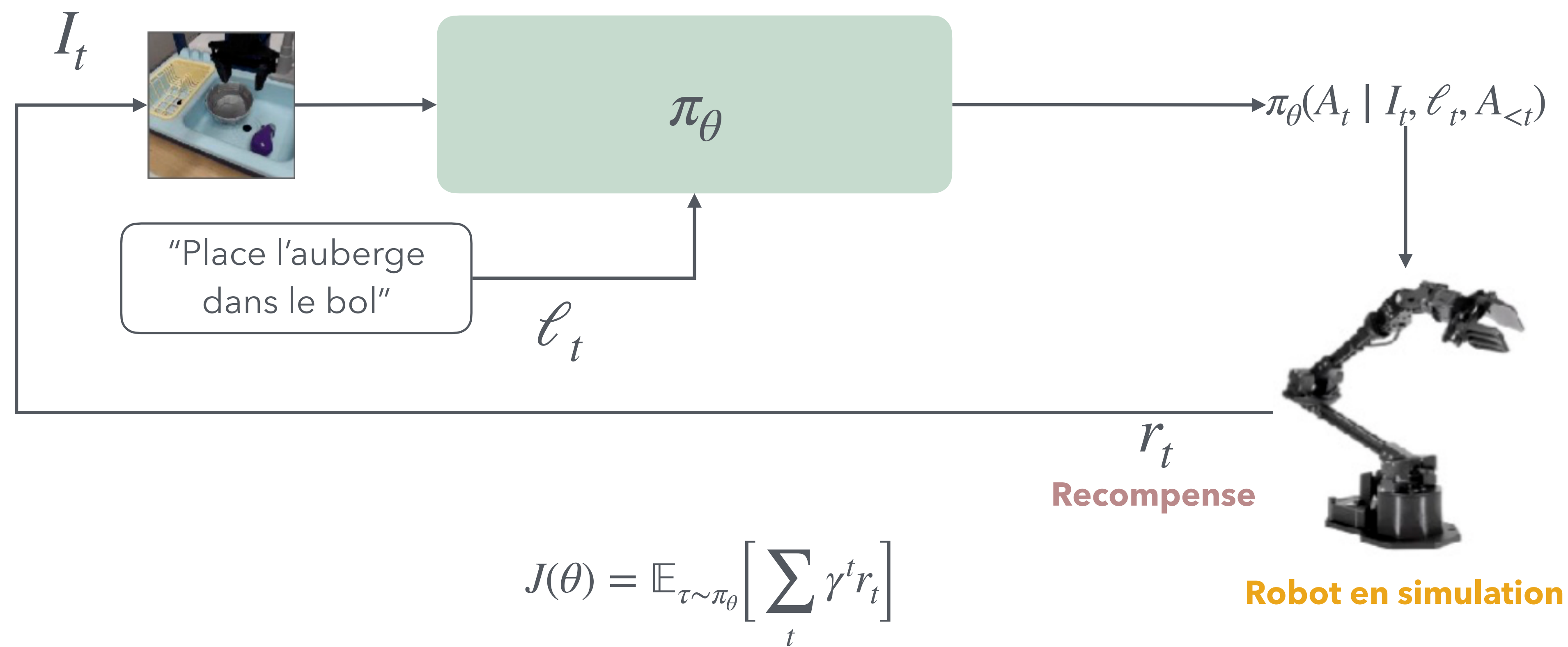
Comment ce genre de modèle est-il entraîné ou pré-entraîné ?



L'Apprentissage par Renforcement (Reinforcement Learning)

Learning and Fine-tuning ?

Apprendre des actions qui ameliore le retour attendu



L'Apprentissage par Renforcement (Reinforcement Learning)

Learning and Fine-tuning ?

Apprendre des actions qui améliore le retour attendu

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_t \gamma^t r_t \right]$$

L'Apprentissage par Renforcement (Reinforcement Learning)

Learning and Fine-tuning ?

Apprendre des actions qui améliore le retour attendu

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_t \gamma^t r_t \right]$$

Iteration 0

Learning bipedal walking in Isaac Sim

Credits: Tokyo Robotics

L'Apprentissage par Renforcement (Reinforcement Learning)

Learning and Fine-tuning ?

Apprendre des actions qui améliore le retour attendu

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_t \gamma^t r_t \right]$$

Iteration 0



Learning bipedal walking in Isaac Sim

Credits: Tokyo Robotics

L'Apprentissage par Renforcement (Reinforcement Learning)

Learning and Fine-tuning ?

Apprendre des actions qui améliore le retour attendu

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_t \gamma^t r_t \right]$$

Iteration 0



Learning bipedal walking in Isaac Sim

Credits: Tokyo Robotics

Tesla Optimus cooking and cleaning

Autres modele VLA et methodes

Actions discretisés

1. Valeurs sont converties en tokens;
2. Méthode: la prédiction se fait en tokens.
3. Les actions de commande sont décodées.

Ex: OpenVLA (<https://openvla.github.io/>), RT-2 (<https://robotics-transformer2.github.io/>)

Actions continues

1. **Distribution** d'actions;
2. Méthode: **Diffusion ou flow matching**;
3. **Génération d'une séquence.**

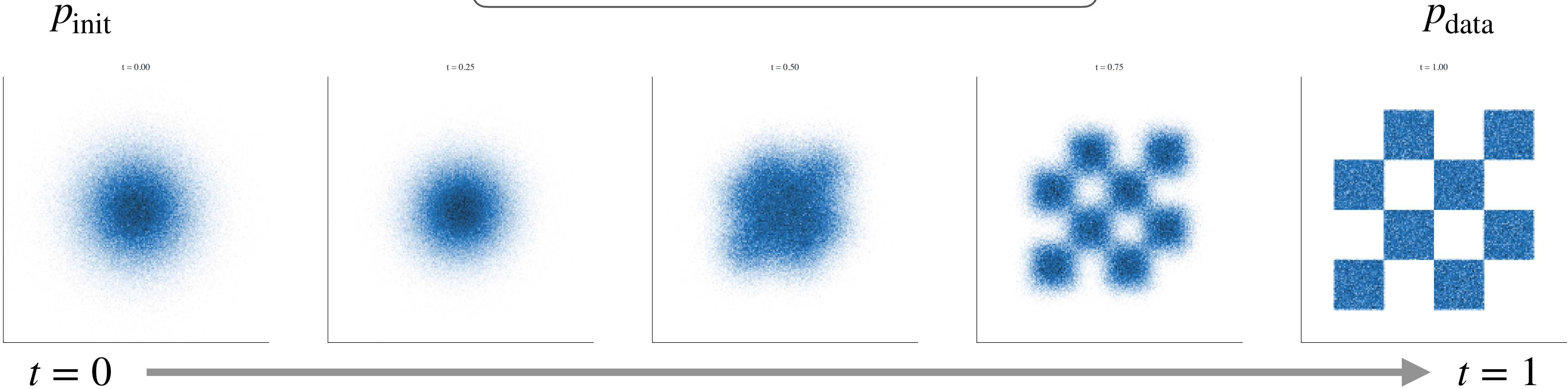
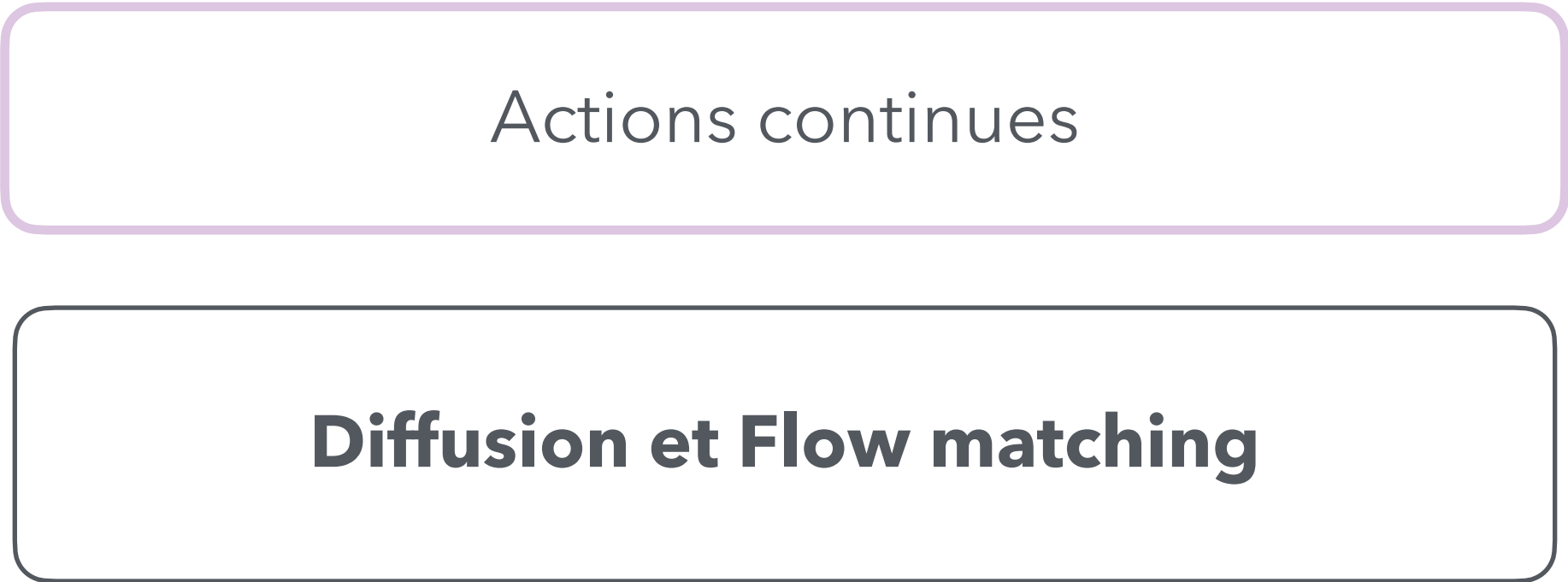
Ex: la famille des π_0 (<https://www.pi.website/blog/pi0>)

Une prochaine fois?

Actions continues

Diffusion et Flow matching

Une prochaine fois?



Reference et Lectures recommandées

- RT-2, Brohan et al., 2023. Comprendre le transfert des connaissances vision-langage vers les actions robotiques (<https://robotics-transformer2.github.io/>).
- OpenVLA, Kim et al., 2024. Étudier une architecture ouverte, son code et son adaptation a de nouvelles tâches (<https://openvla.github.io/>).
- Diffusion Policy, Chi et al., 2023. Comprendre la génération de séquences d'actions par diffusion. (<https://diffusion-policy.cs.columbia.edu/>).
- π_0 , Black et al., 2024. Explorer un VLA qui génère des actions continues grâce au flow matching. (<https://arxiv.org/abs/2410.24164>).
- $\pi_{0.6}$, Physical Intelligence, 2025. Découvrir comment améliorer un VLA grâce à l'expérience et à l'apprentissage par renforcement. (<https://arxiv.org/abs/2511.14759>).

Thank you!

Questions?

Alex Zongo

Department of Mechanical and Aerospace Engineering
George Washington University, Washington, DC, USA.

Contact: a.zongo@gwu.edu



alexzongo.com