DEEP LEARNING ROBOTICS

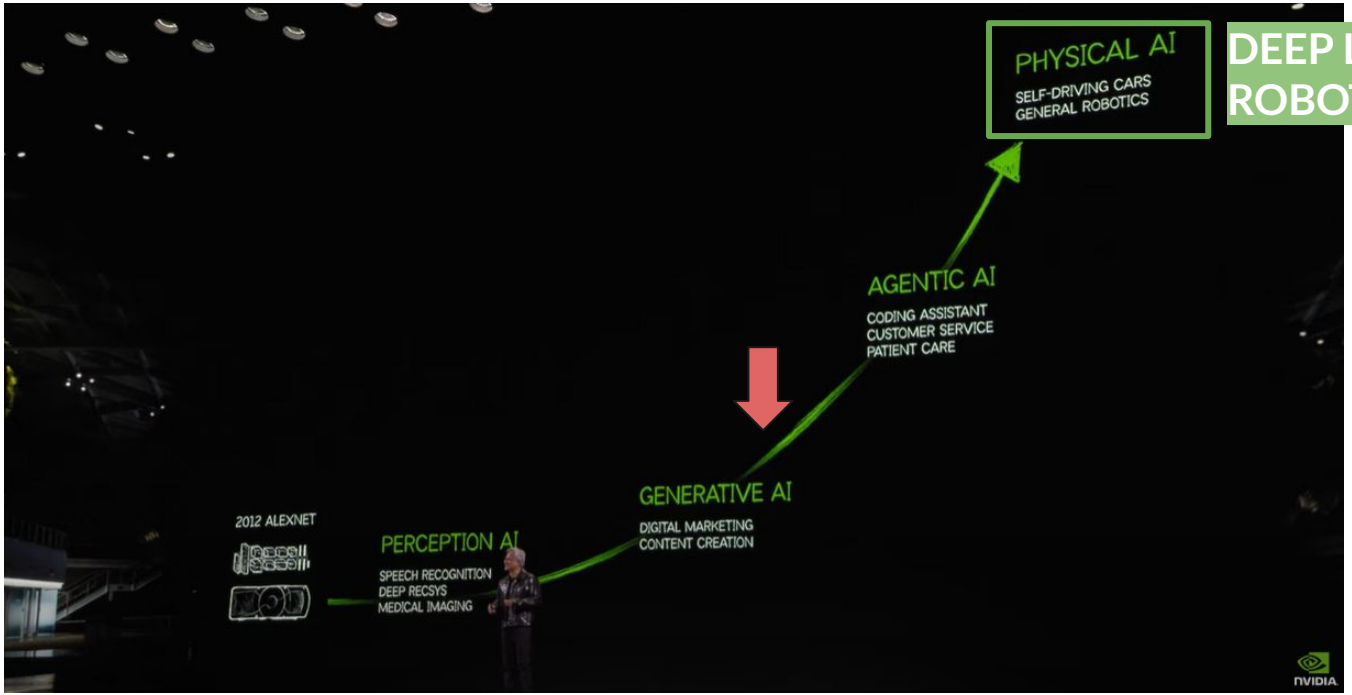
Course of Intelligent Robotics 2024-2025

Alberto Bacchin

alberto.bacchin.1@phd.unipd.it

bacchinalb@dei.unipd.it

A new technology breakout?



Jen-Hsun Huang at CES 2025

Prerequisites

- Deep Learning
- Robotics
- Computer Vision and/or 3D Data Processing

Outlines

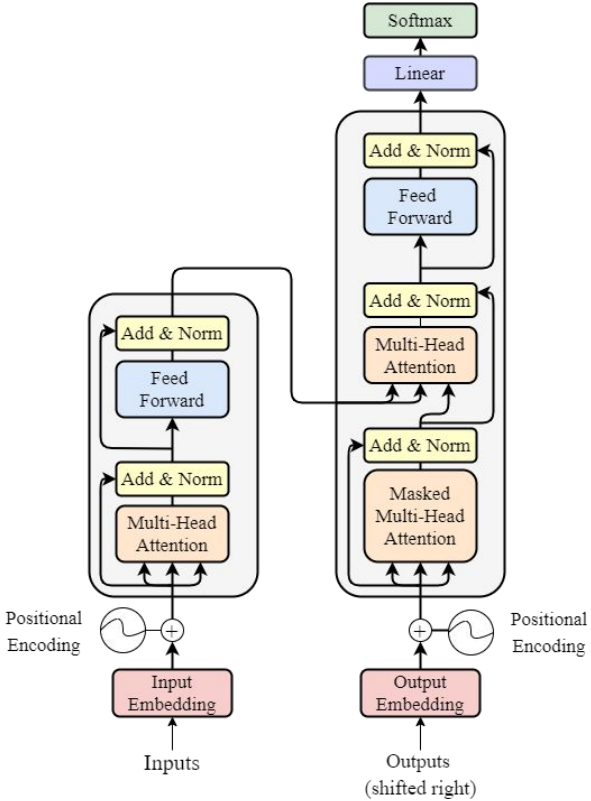
- The Role of Transformers
- Deep Learning applied to Robotics
- Our Research and Thesis

DISCLAIMER: this is just an introduction to the topic, that is **large** and **rapidly changing**.

The Role of Transformers

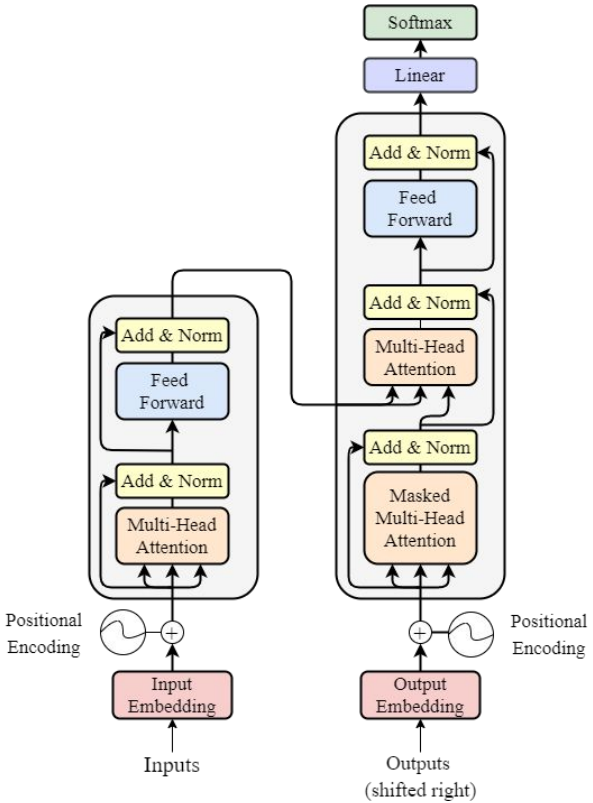
Transformer Architecture

Transformer Architecture



Vaswani, A. "**Attention** is all you need." *Advances in Neural Information Processing Systems* (2017).

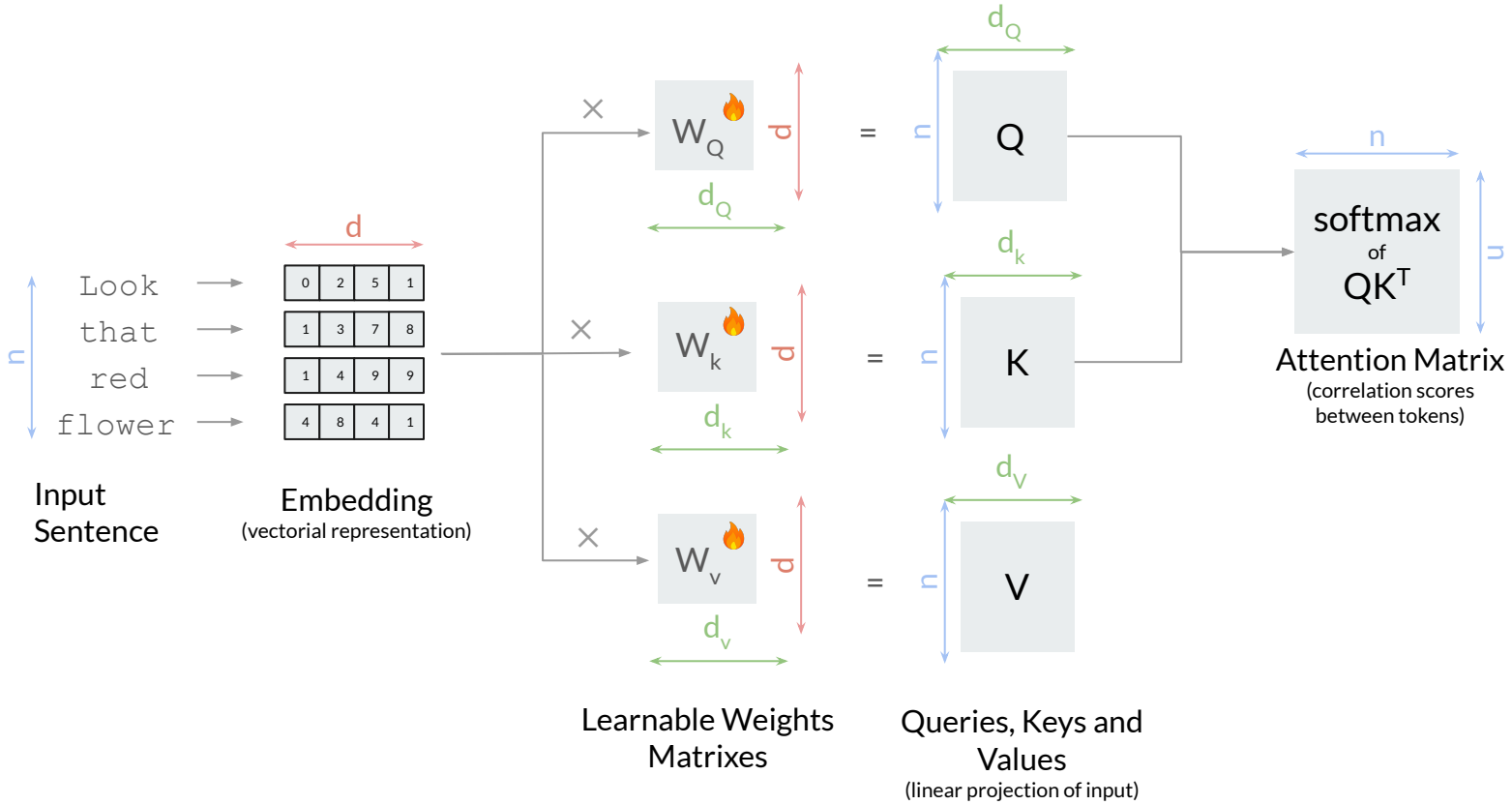
Transformer Architecture



- Developed for NLP (language translation)
- Process **sequential** inputs (a list of tokens…)
- Key Ideas:
  a. **Transform** the input in a different representation…
  b. …that accounts of **contextual information**

Vaswani, A. "**Attention** is all you need." *Advances in Neural Information Processing Systems* (2017).

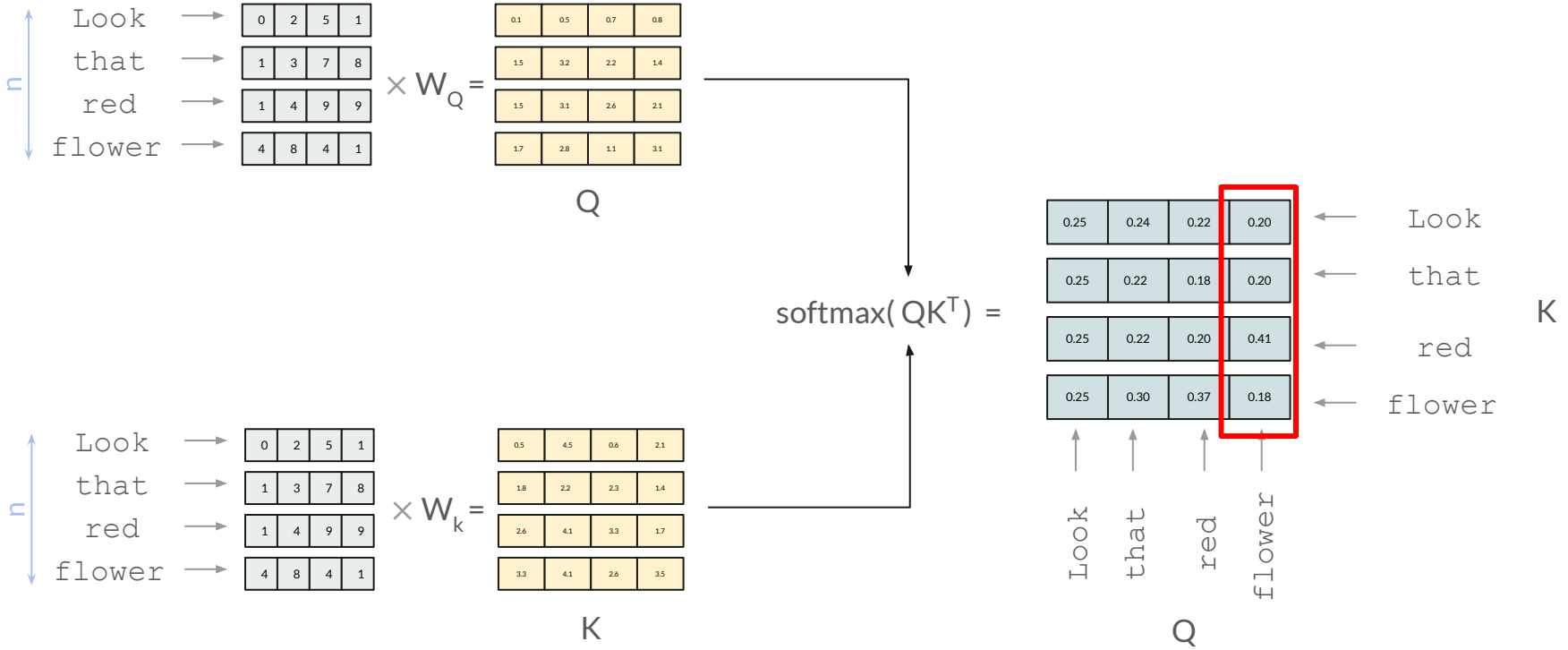
The Attention Mechanism (in NLP)



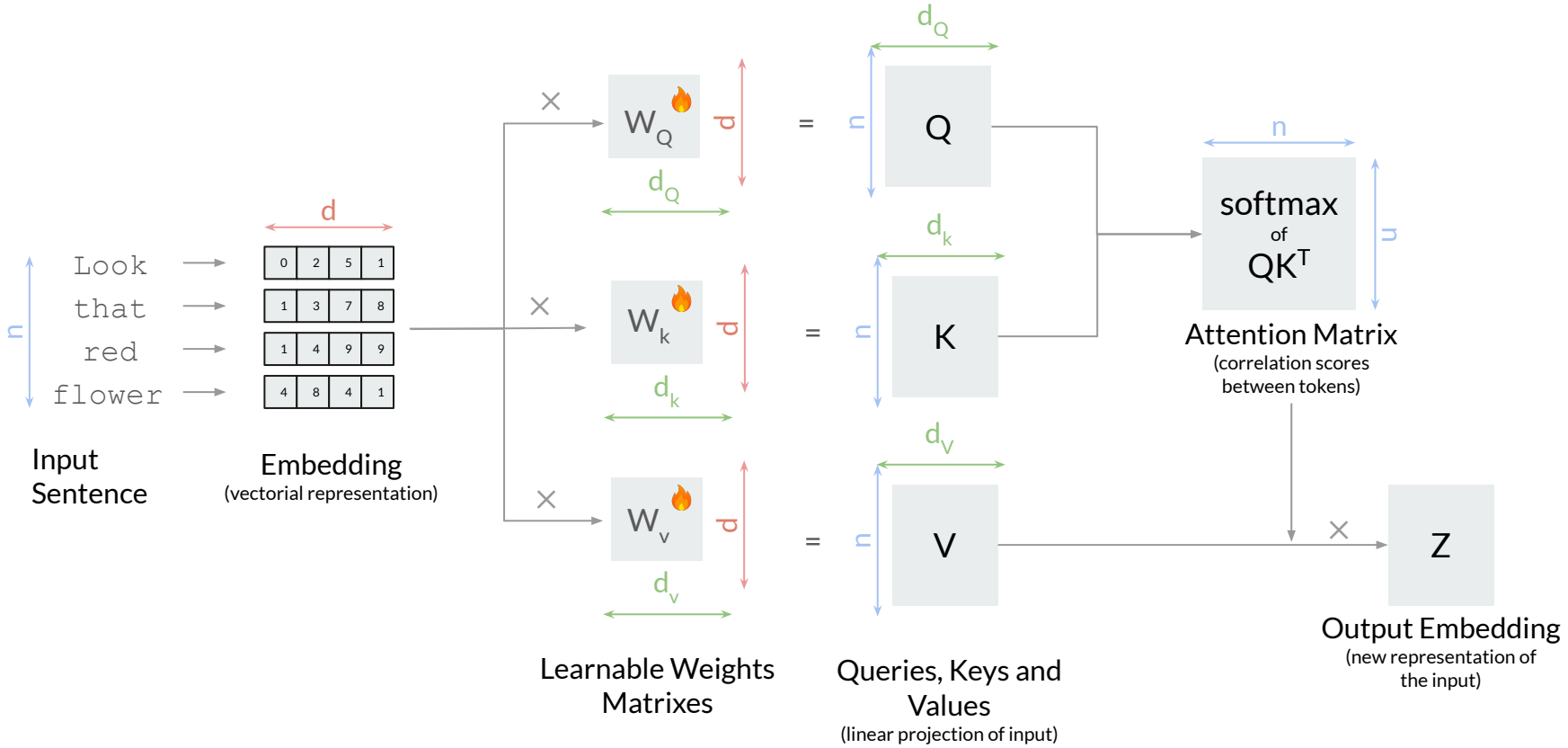
Attention Matrix

Attention Matrix is representing correlation relationships between Queries and Keys

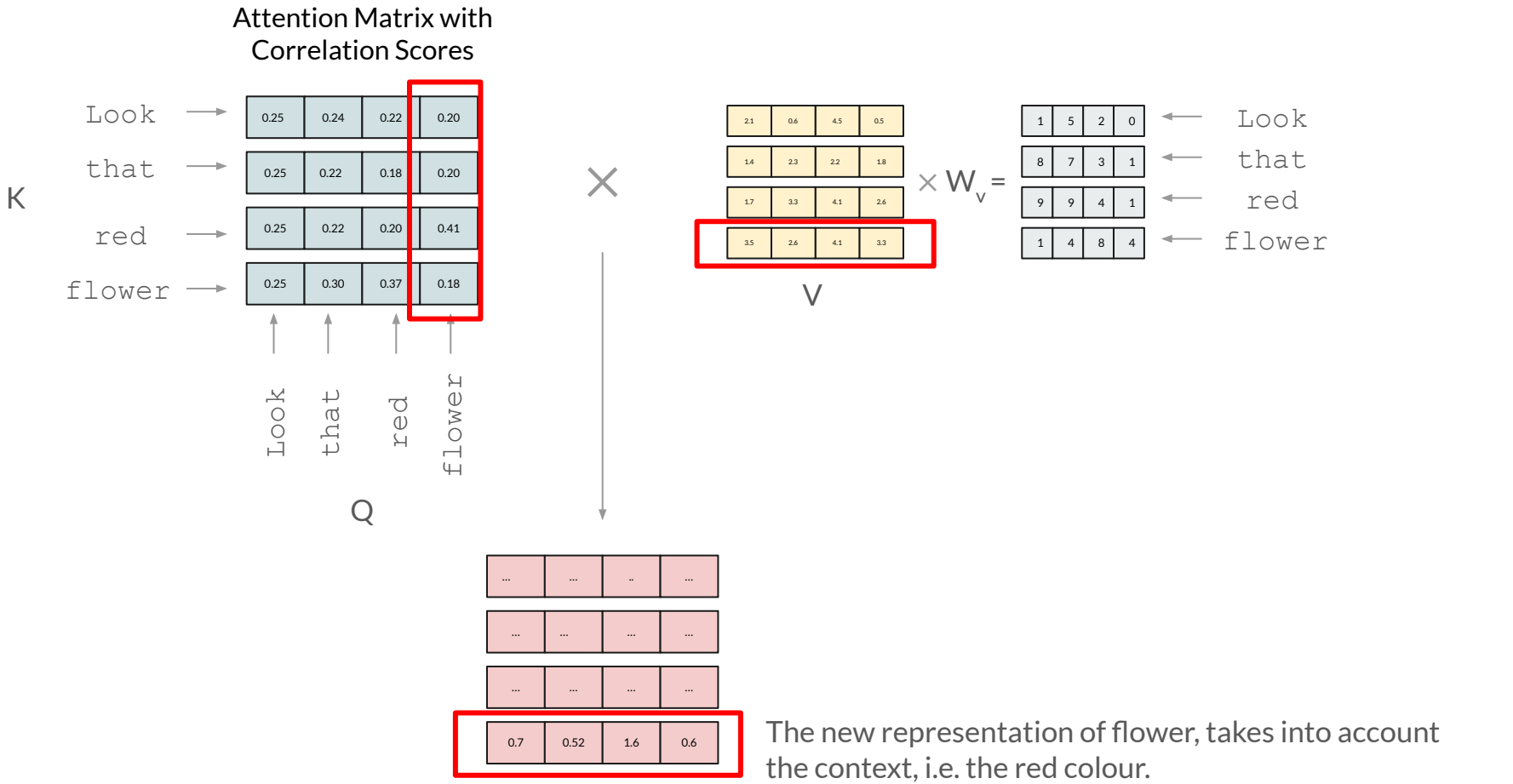
- Queries can be seen as “questions” → is there an adjective closed me? (me=a certain token)
- Keys can be seen as possible “answers”
- Attention scores can be seen as a “measure of correctness” of a key to a query



The Attention Mechanism (in NLP)



The Output Embedding



The new representation of flower, takes into account the context, i.e. the red colour.

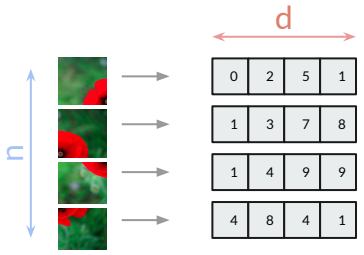
Attention Beyond NLP

Transformers process sequences of tokens
→ reshape images as sequences

Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

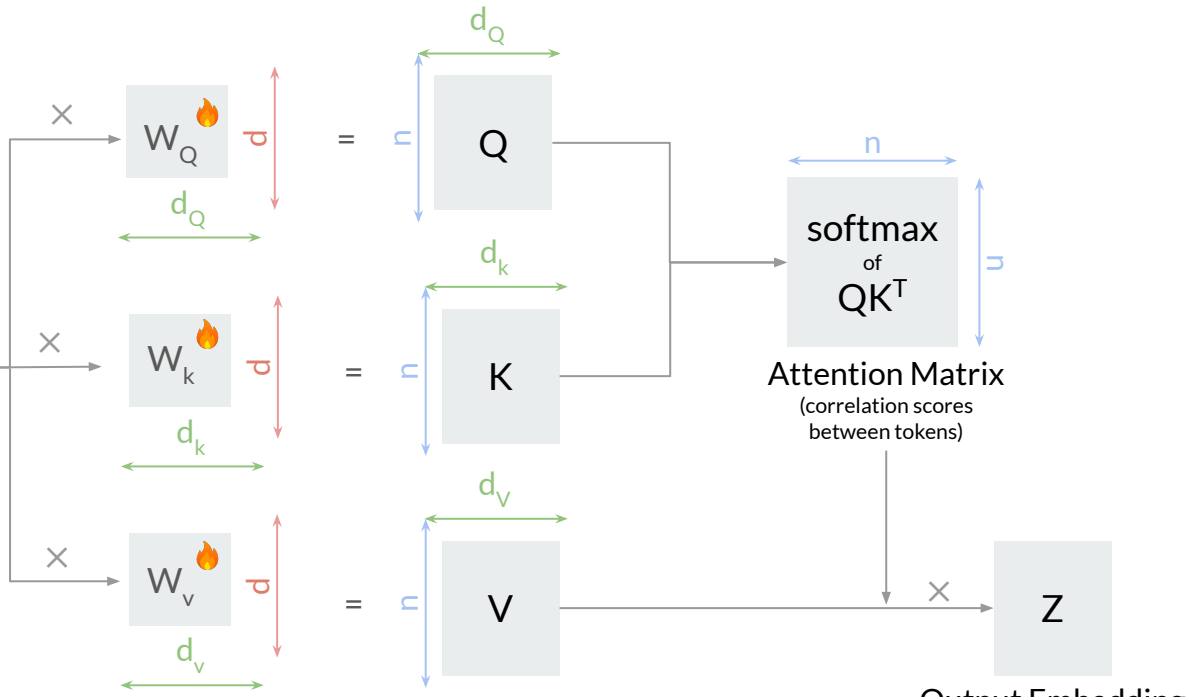
Attention Beyond NLP

Attention Beyond NLP



Input Image

Embedding
(vectorial representation)



Learnable Weights Matrixes

Queries, Keys and Values
(linear projection of input)

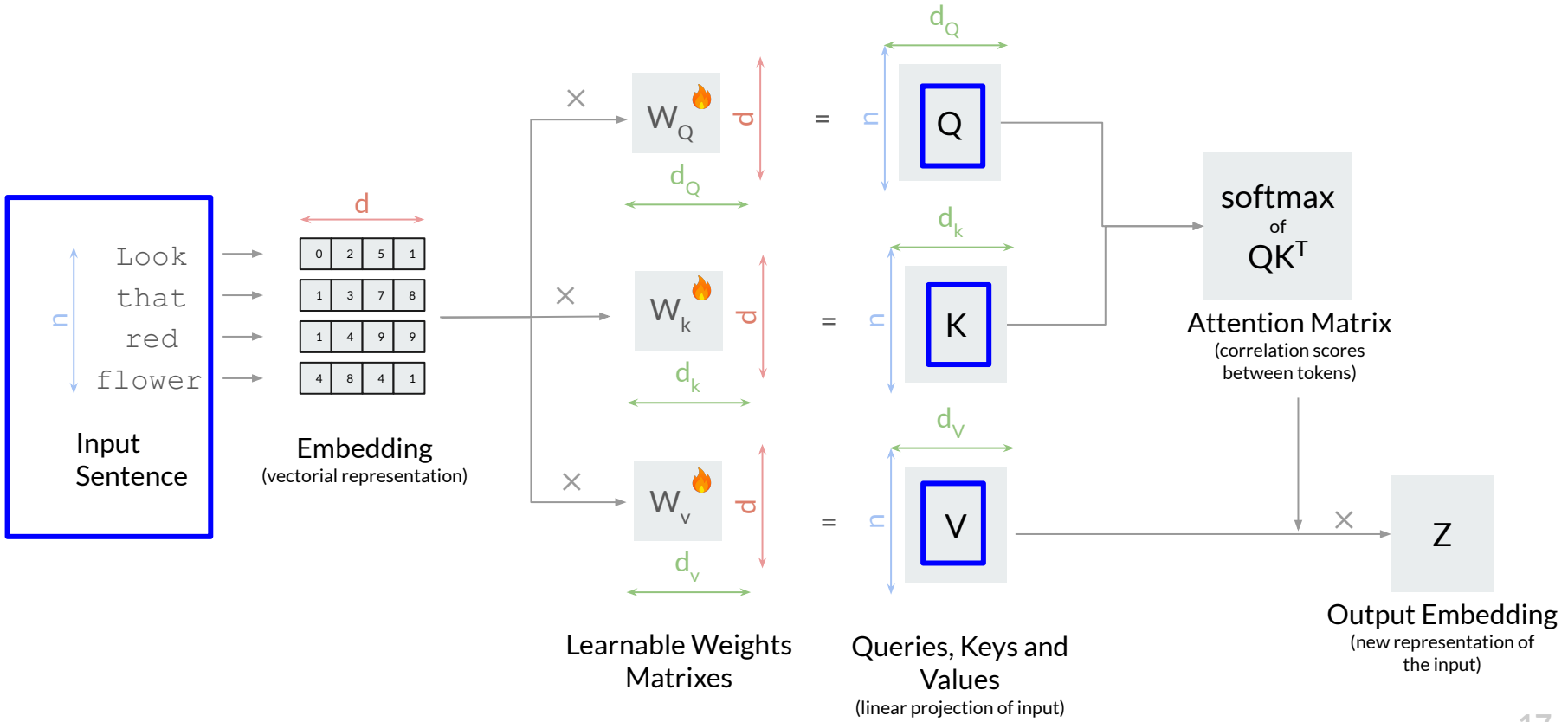
We can use attention for different kinds of inputs!

Self-Attention and Cross-Attention

Why is this useful for robotics?

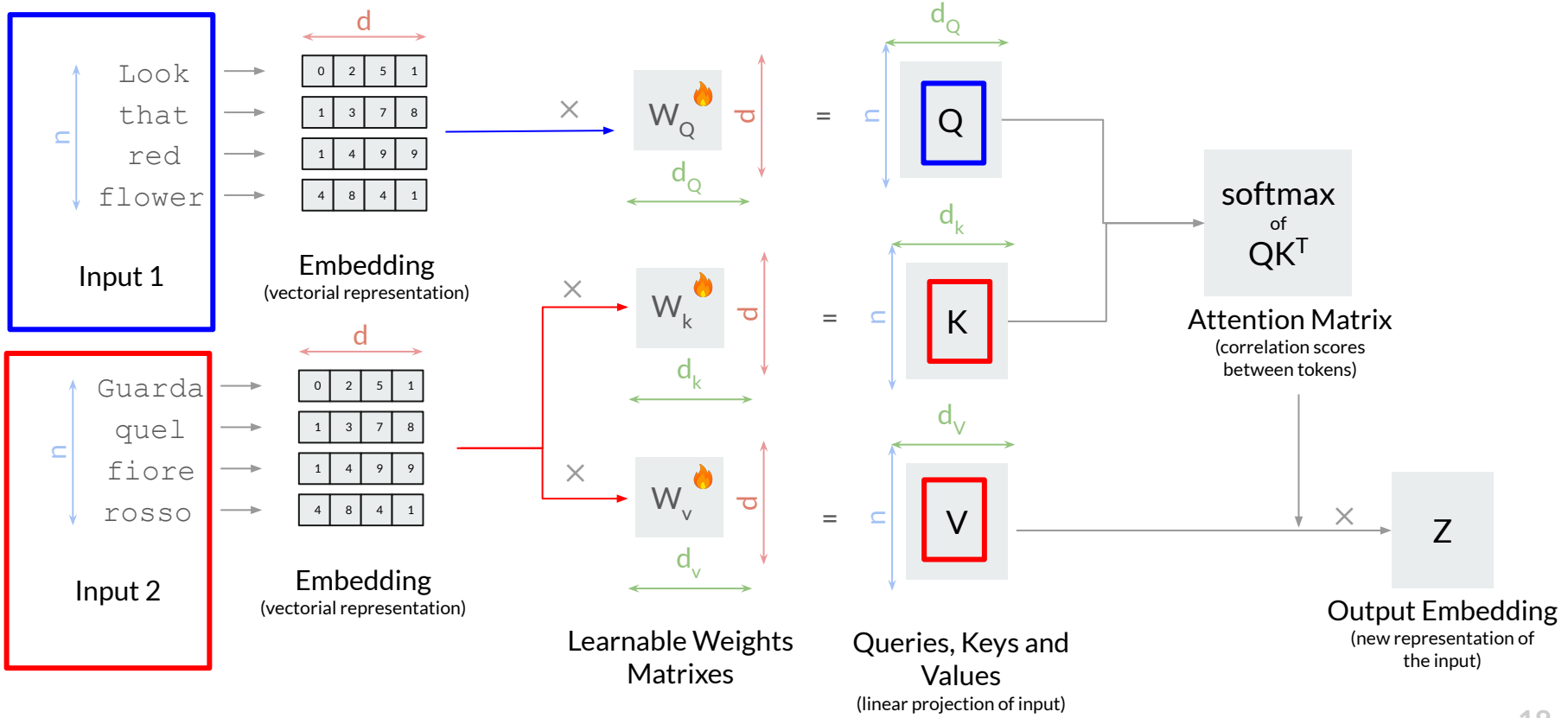
Self-Attention and Cross-Attention

SELF-ATTENTION



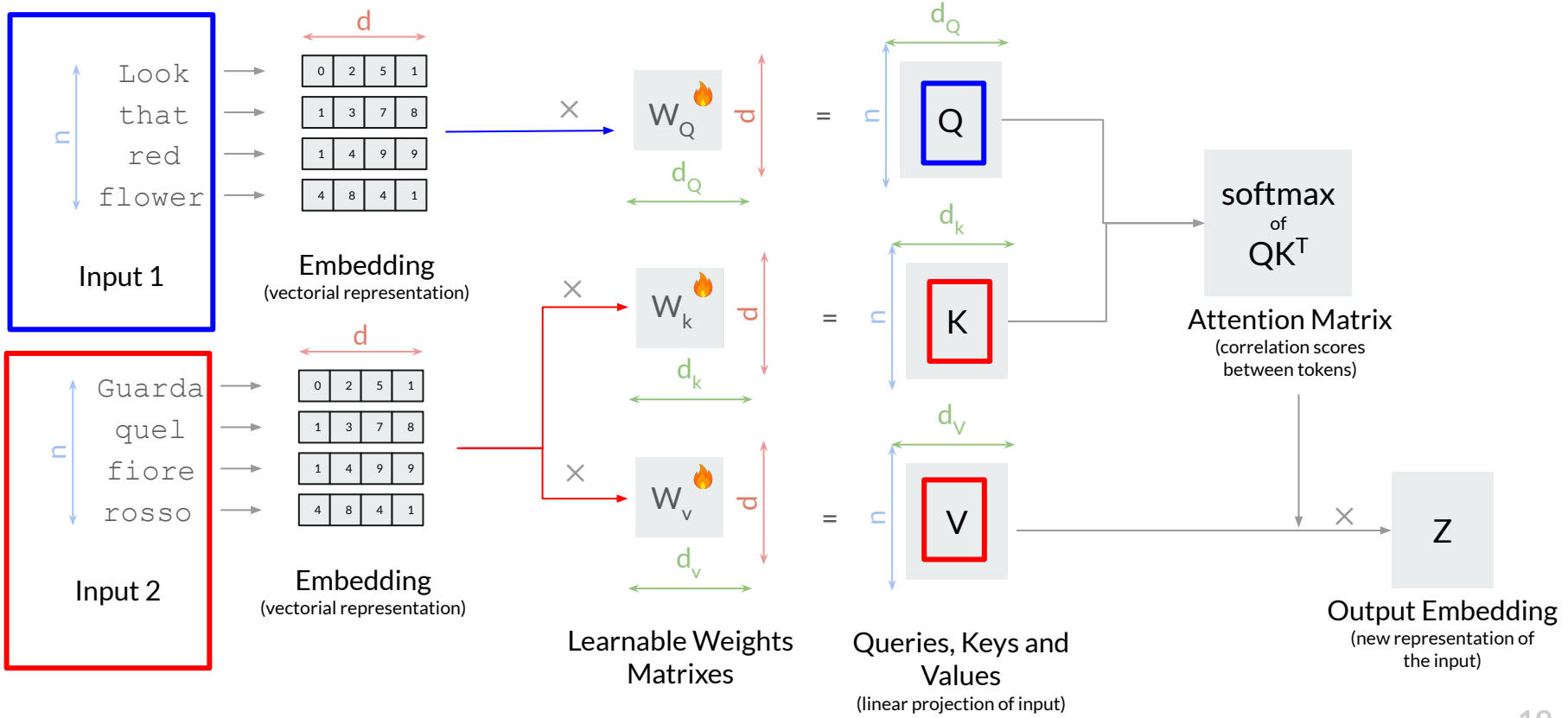
Self-Attention and Cross-Attention

CROSS-ATTENTION



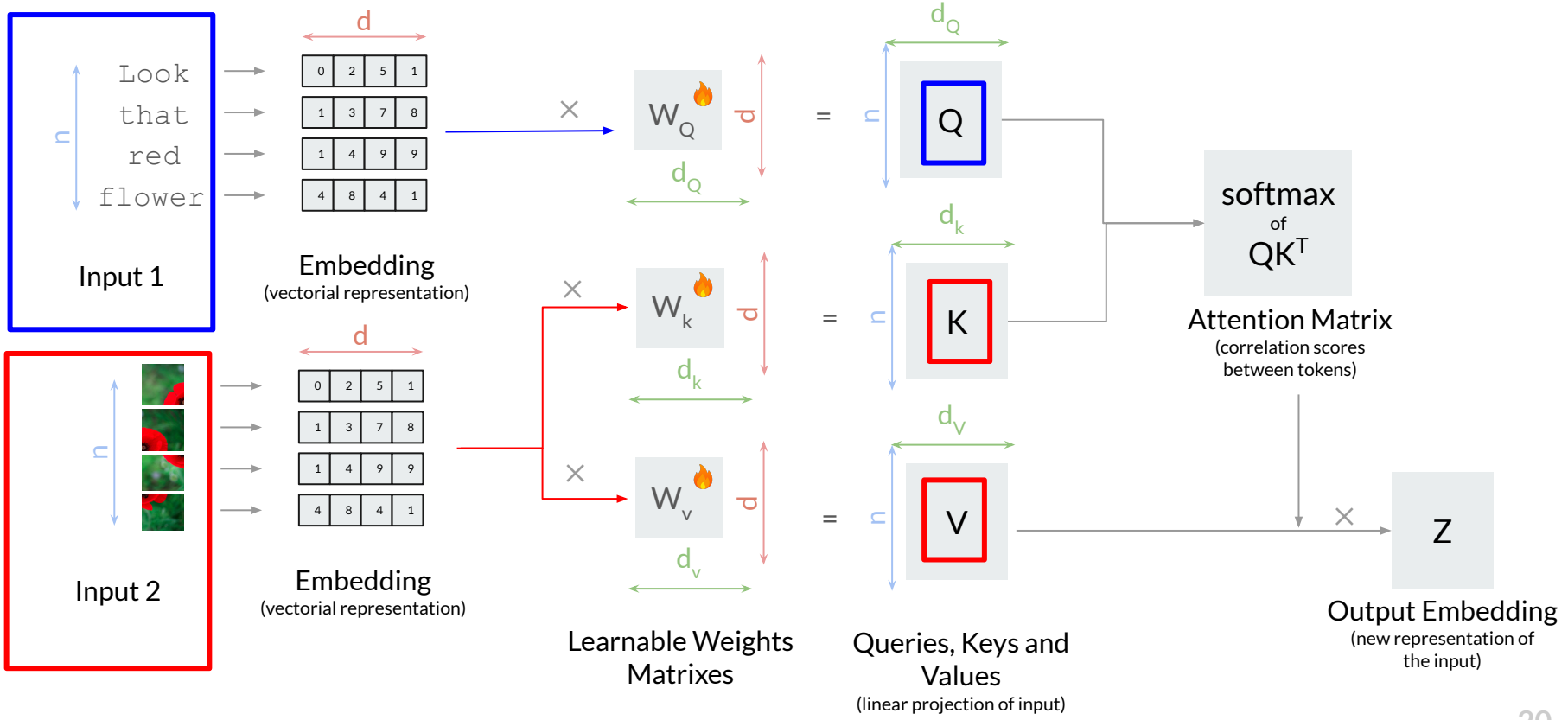
Self-Attention and Cross-Attention

You get a new representation Z of input 2 that depends on the correlation between input1 and input2



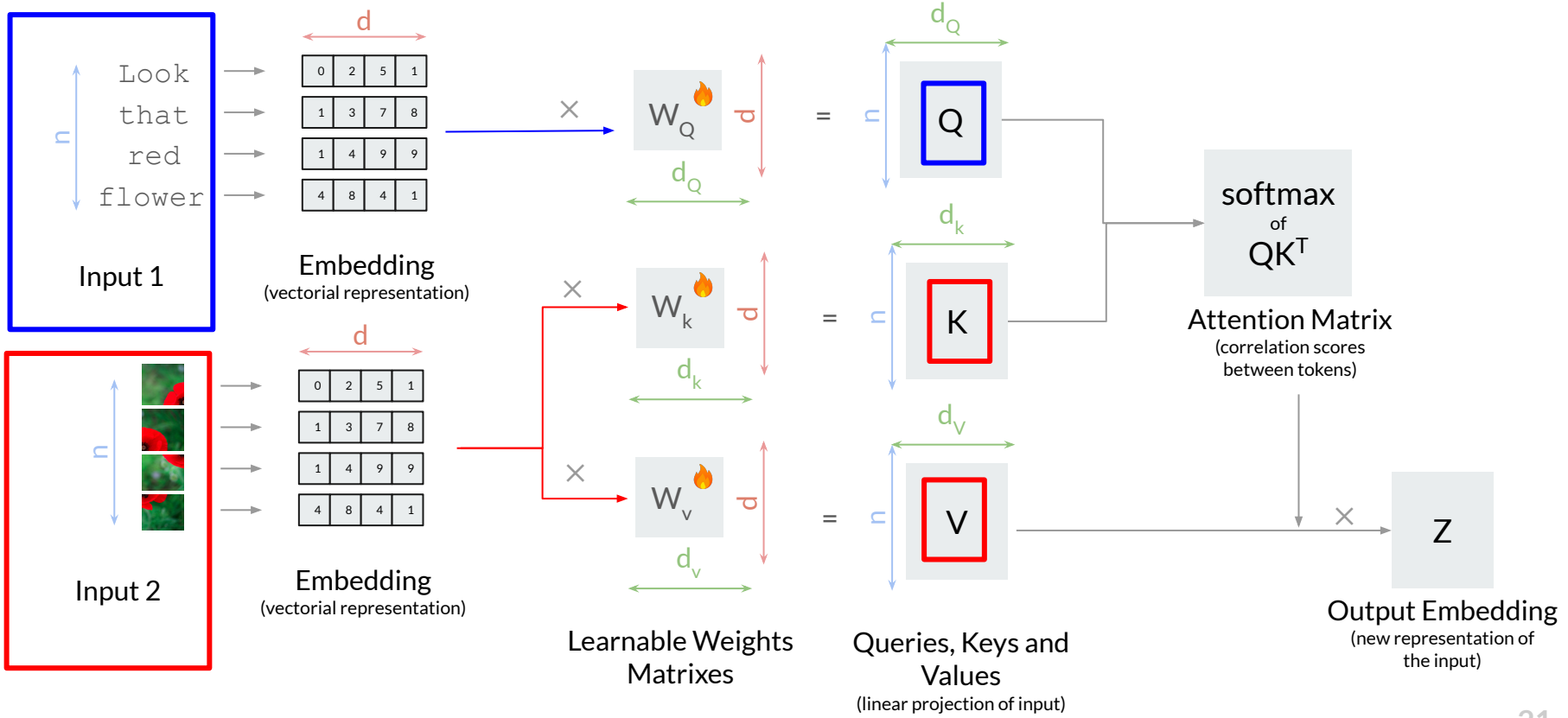
Self-Attention and Cross-Attention

This allows to **mix different inputs.**

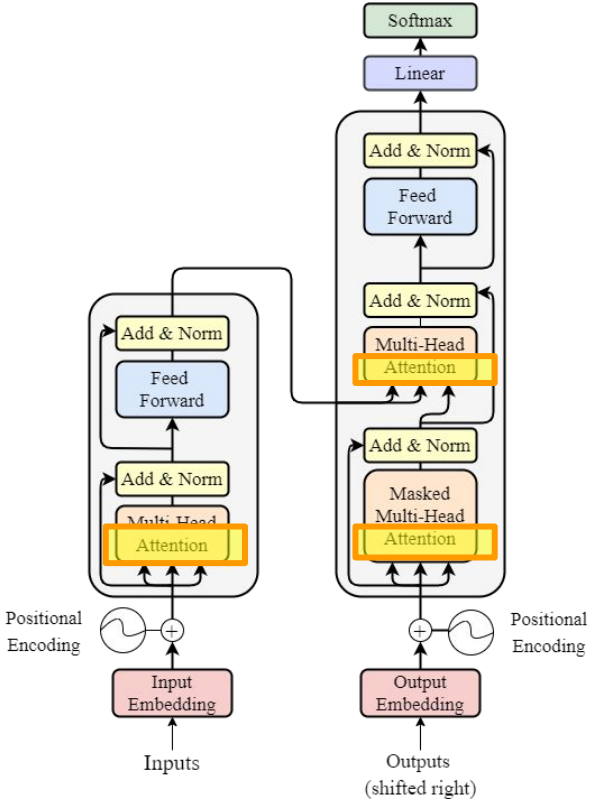


Self-Attention and Cross-Attention

This allows to **mix different inputs. → USEFUL FOR ROBOTS**

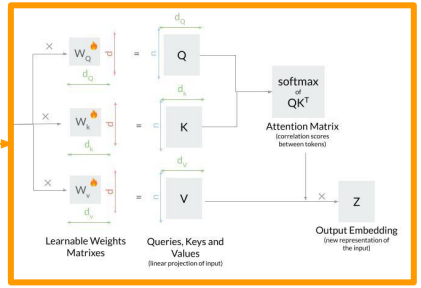


Transformer Architecture



Vaswani, A. "**Attention** is all you need." *Advances in Neural Information Processing Systems* (2017).

Single-Head Attention



Z_1 is like a transformation to apply to the original embedding to include contextual informations

Z_1 +

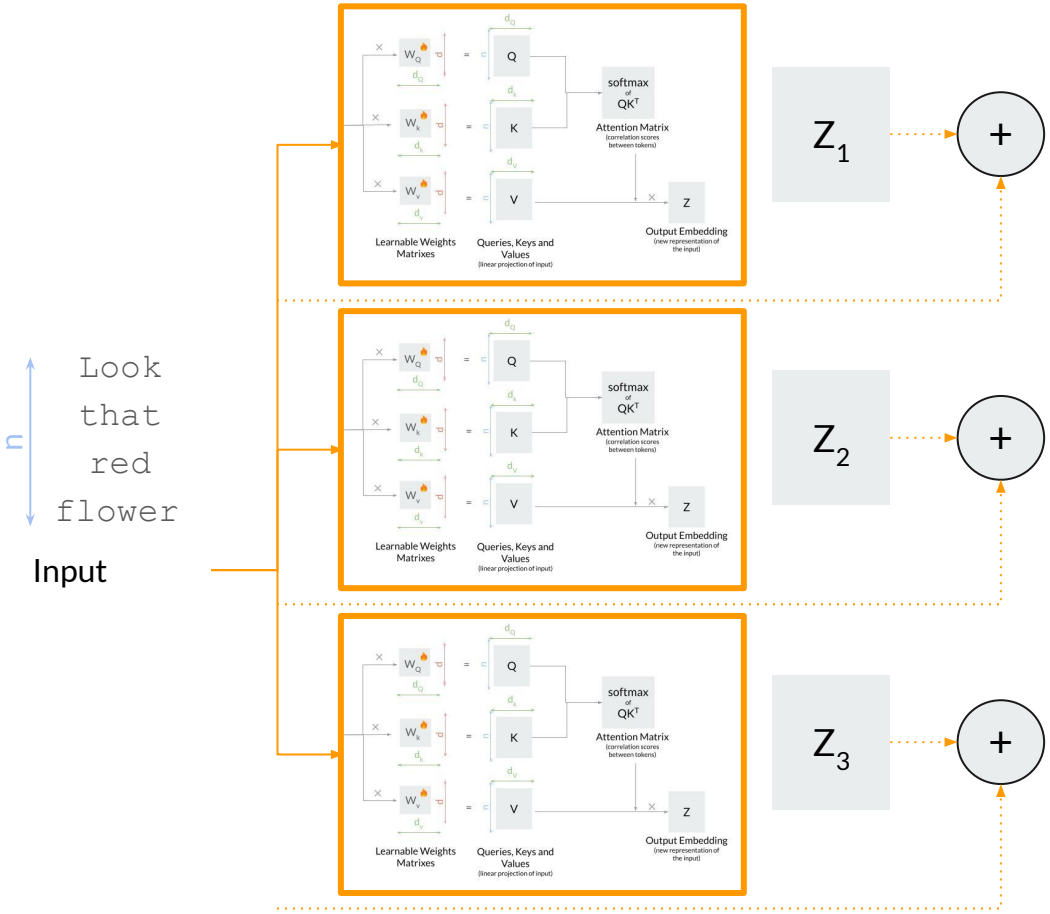
Residual Connections

n

Look
that
red
flower

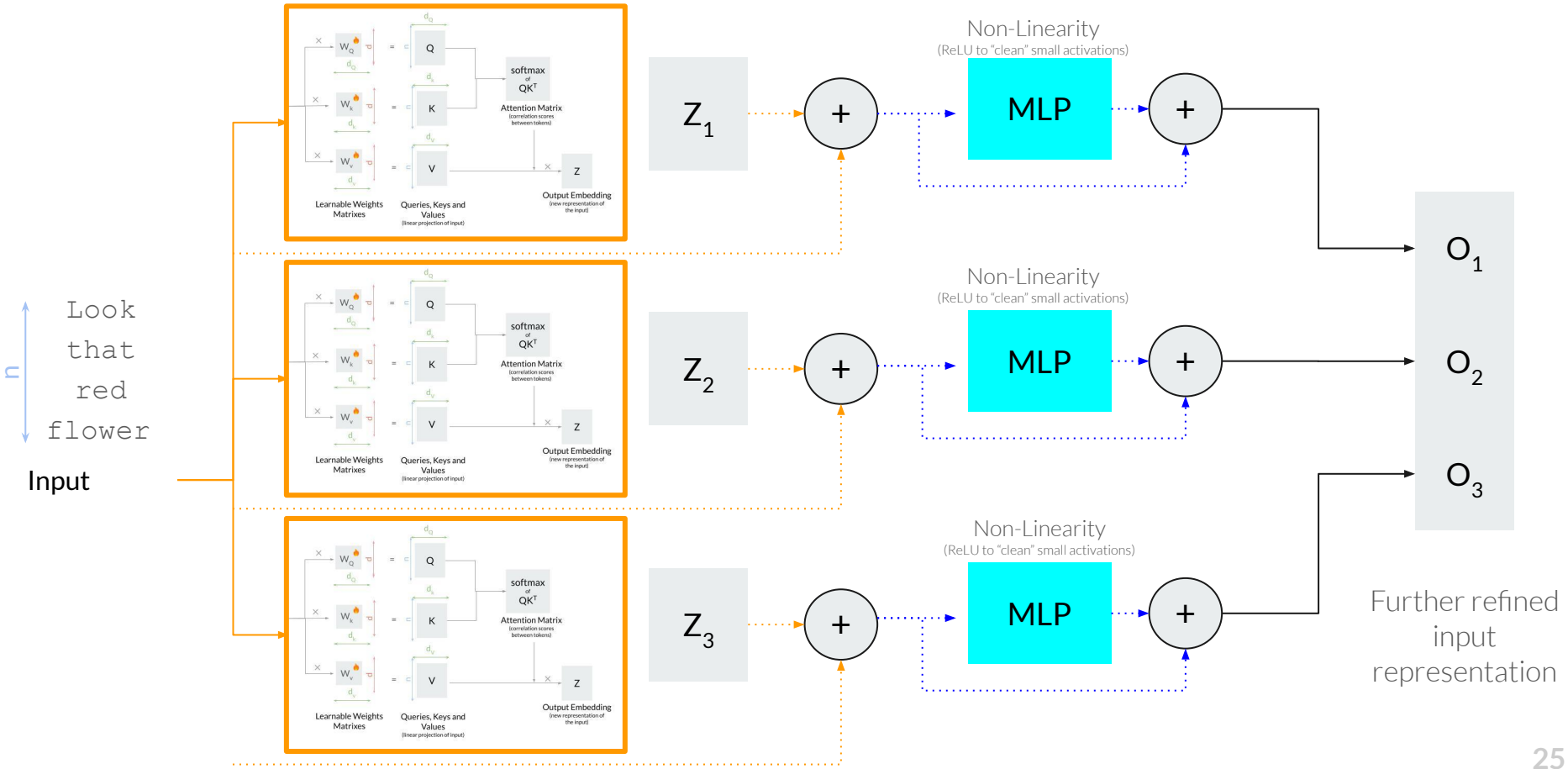
Input

Multi-Head Attention

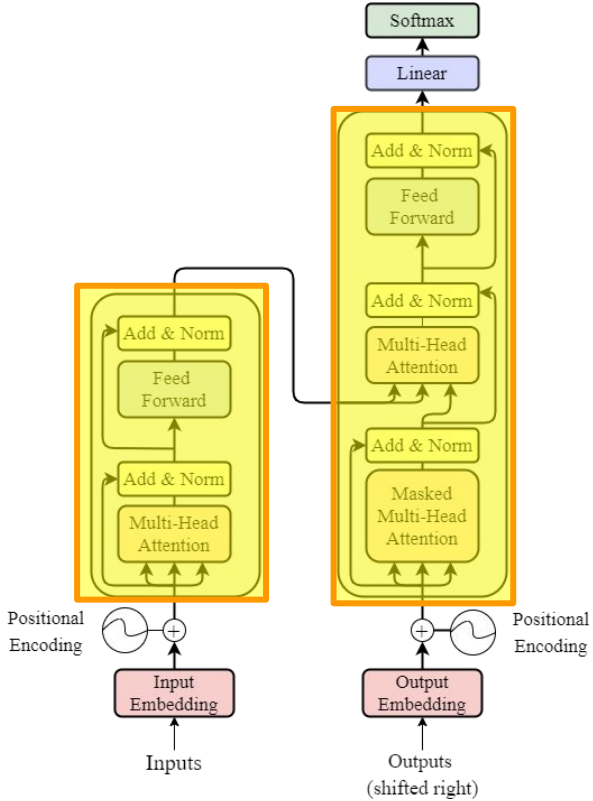


Each Z_i is a different transformation, so you get many representation of your input that embeds different information (Q, K, V are different in each head).

Transformer Block

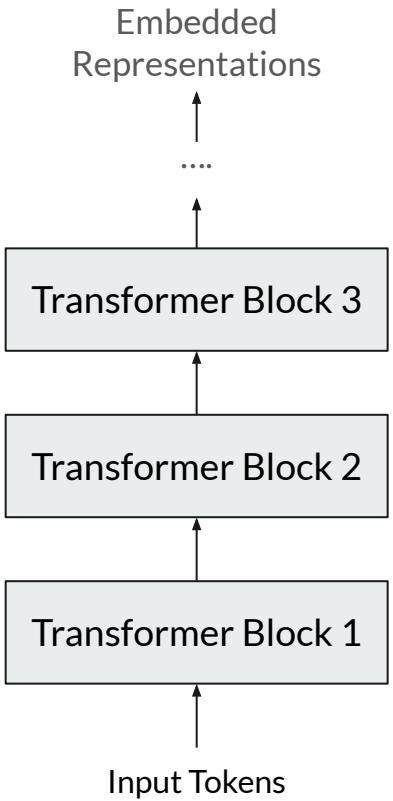


Transformer Block



Vaswani, A. "**Attention** is all you need." *Advances in Neural Information Processing Systems* (2017).

The Transformer



ENCODER TRANSFORMER

Start from raw input tokens and maps into embeddings

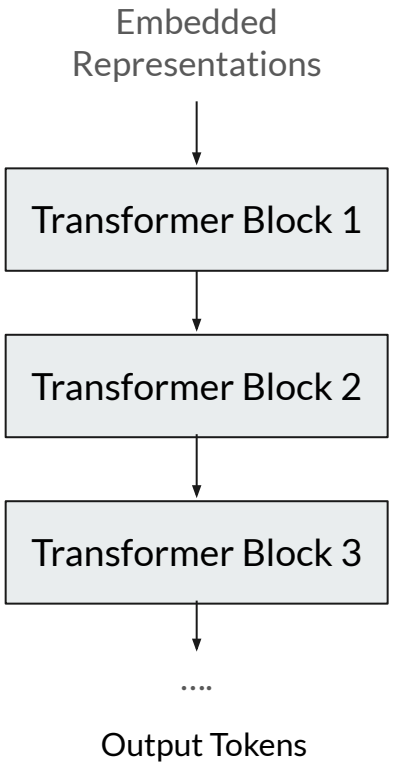
Embeddings can be used to make predictions, store knowledge, etc..

The Transformer

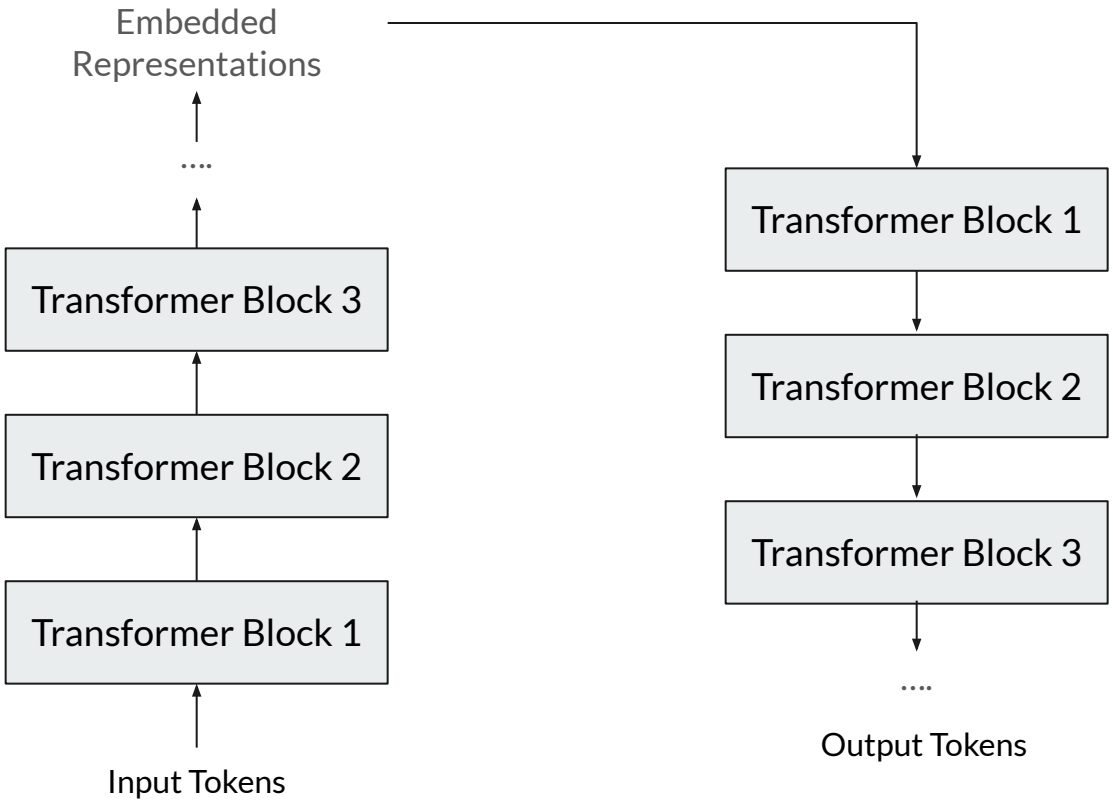
DECODER TRANSFORMER

Start from embeddings and maps into output tokens

Used for generation, like in ChatBots



The Transformer

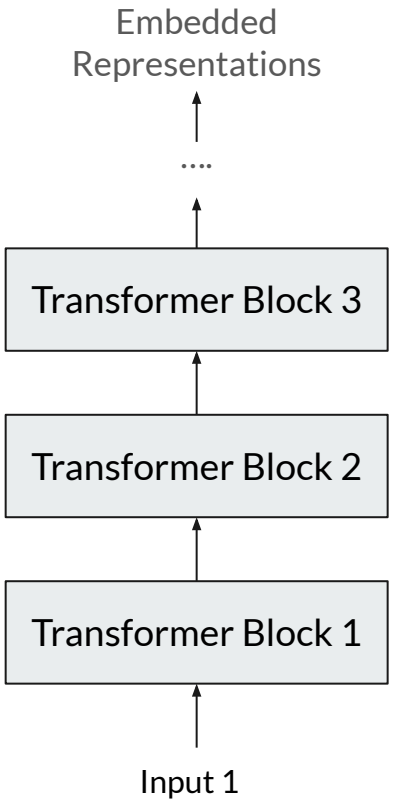


ENCODER-DECODER TRANSFORMER

Start from raw input tokens (encoder) and maps into output tokens (decoder)

Used for end-to-end application, like machine translation

The Transformer



Characteristics

- Good in learning **long-term dependencies** in data
- **Scales** with limited issues…
- … but it needs **a lot of data**
- Good in processing **multi-modal data**

Deep Learning Applied to Robotics

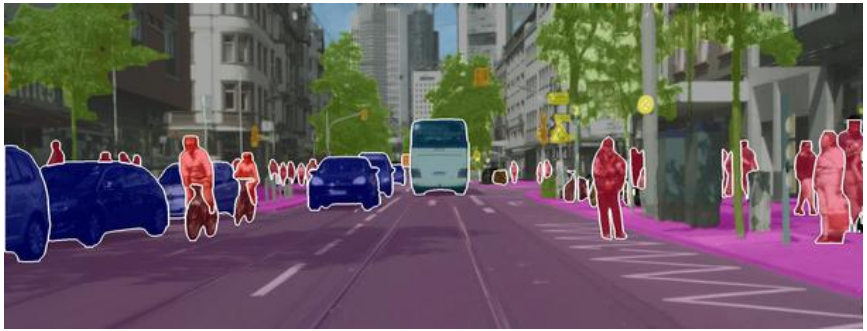
DL and Robotics

- Perception
- Planning
- Sense-Plan-Act All in One (AiO)

N.B. I will present you high-level ideas of **cutting-edge research**.

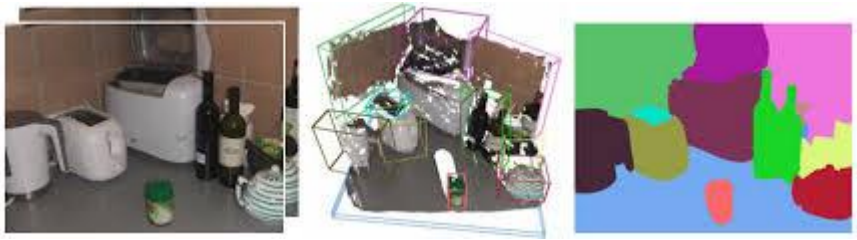
I invite you to **reason on what you have done for your assignment** to understand **how complex and amazing are the methods** I am going to show you.

Perception | Scene Understanding



- Object Detection
- Semantic Segmentation
- Instance Segmentation
- Panoptic Segmentation

→ Computer Vision



- Point Cloud Segmentations
- 6D Pose Estimation

→ 3D Data Processing

- Multi-Object Tracking

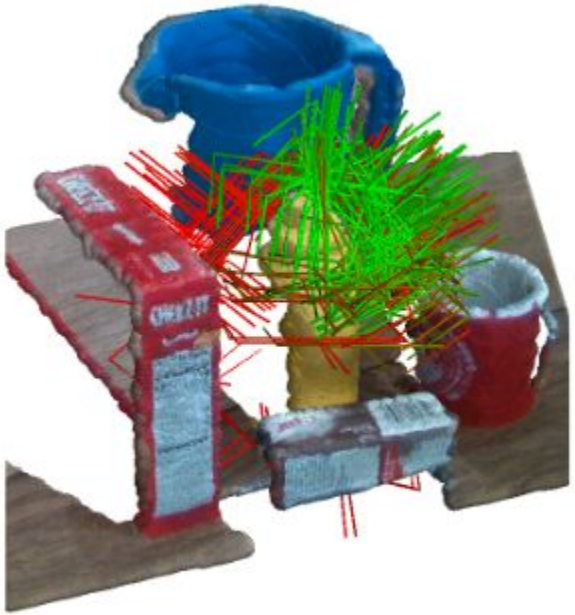
→ Computer Vision

Perception | Robotic Grasping

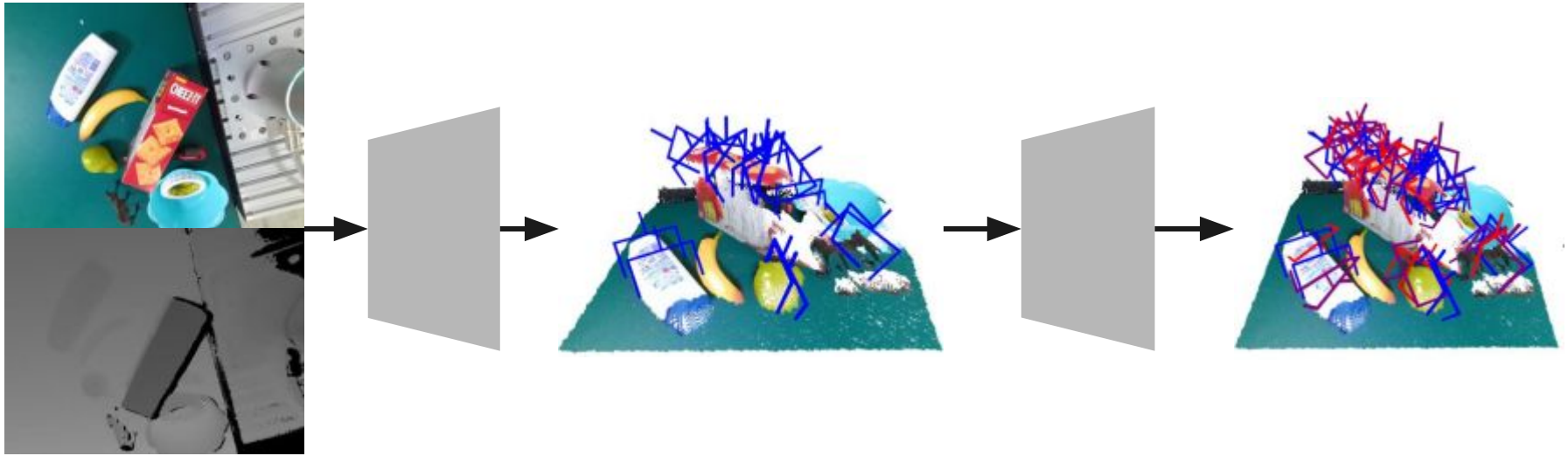
The task of prediction a suitable grasping pose to pick an object

Suitable means feasible (no collisions) and successful (the robot can lift the object).

without Apriltag :(



Perception | Robotic Grasping

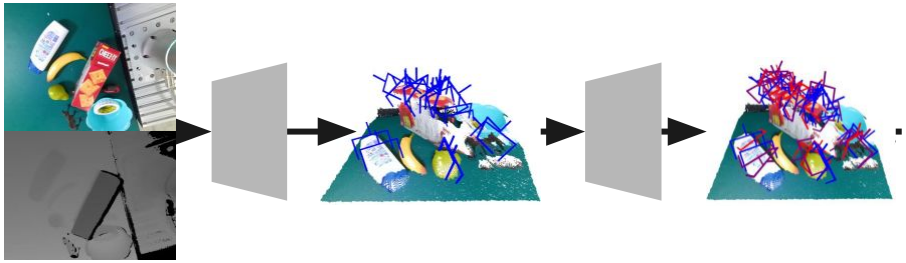


Scene Description
(RGB, RGB-D, point clouds)

Extract Candidate Grasping Poses

Score Candidates

Perception | Robotic Grasping

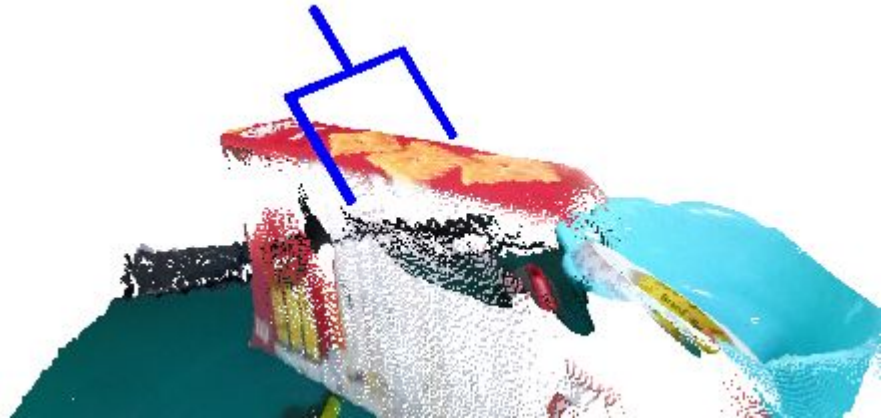


Scene Description
(RGB, RGB-D, point clouds)

Extract Candidate Grasping Poses

Score Candidates

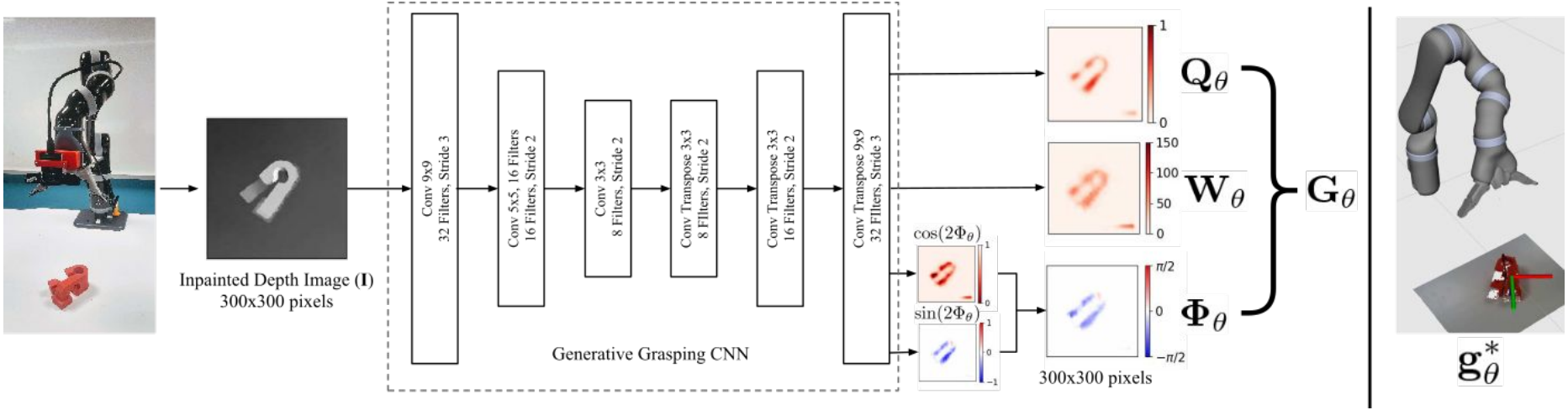
Best Grasping Pose Prediction



Perception | Robotic Grasping

Grasping Pose:
- Quality Score
- Geometric Information (center position, orientation, aperture)

Heatmap-based Robotic Grasping

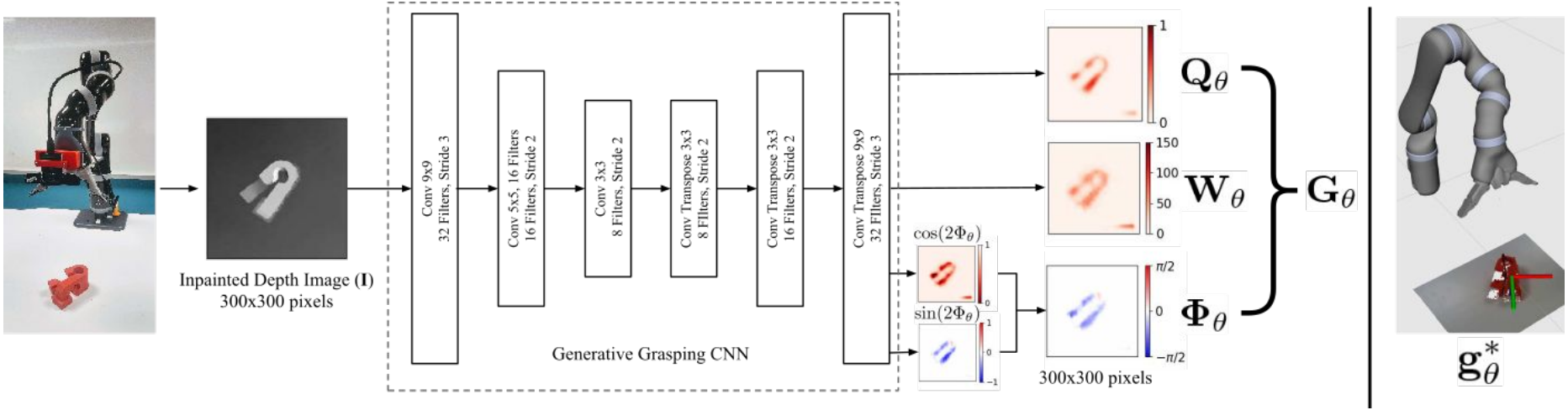


Input Image is mapped into 3 heatmap representations. Heatmaps encode for each pixel a quality score (Q), an orientation (Φ) and a gripper aperture (W).

Perception | Robotic Grasping

Grasping Pose:
- Quality Score
- Geometric Information (center position, orientation, aperture)

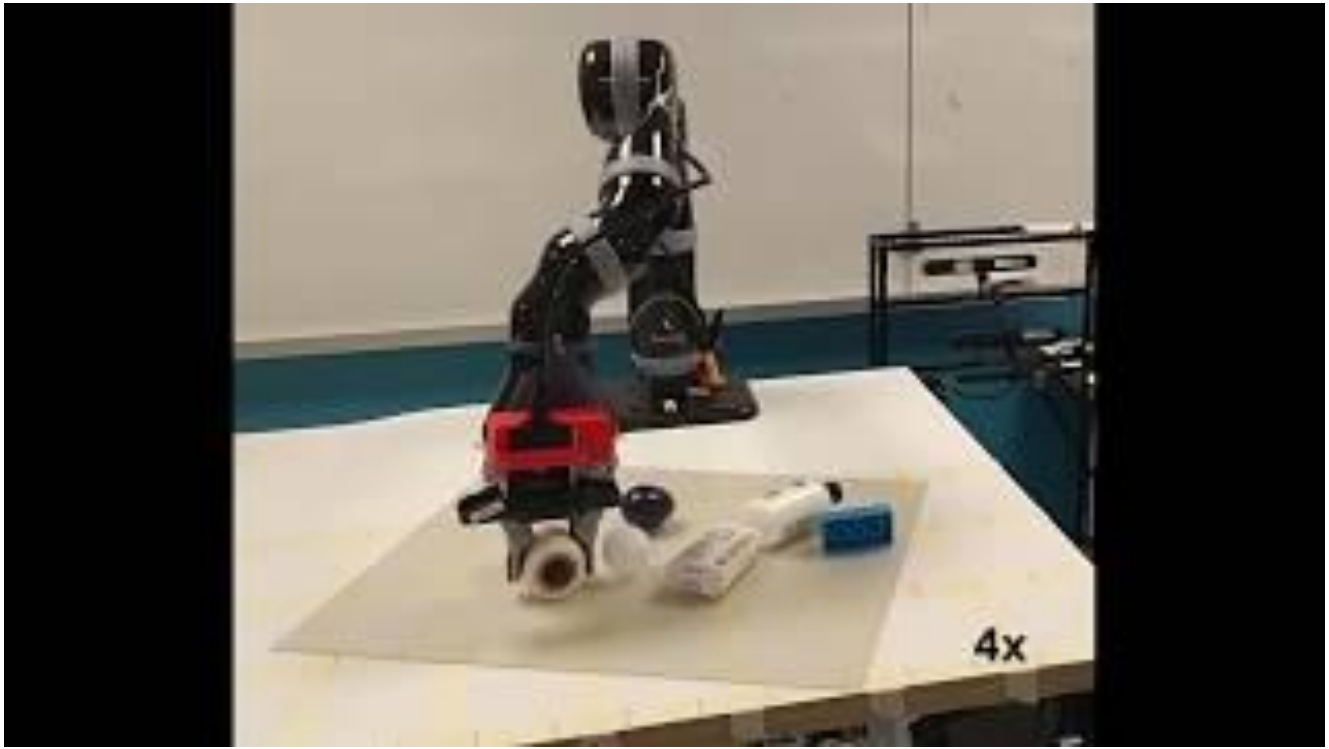
Heatmap-based Robotic Grasping



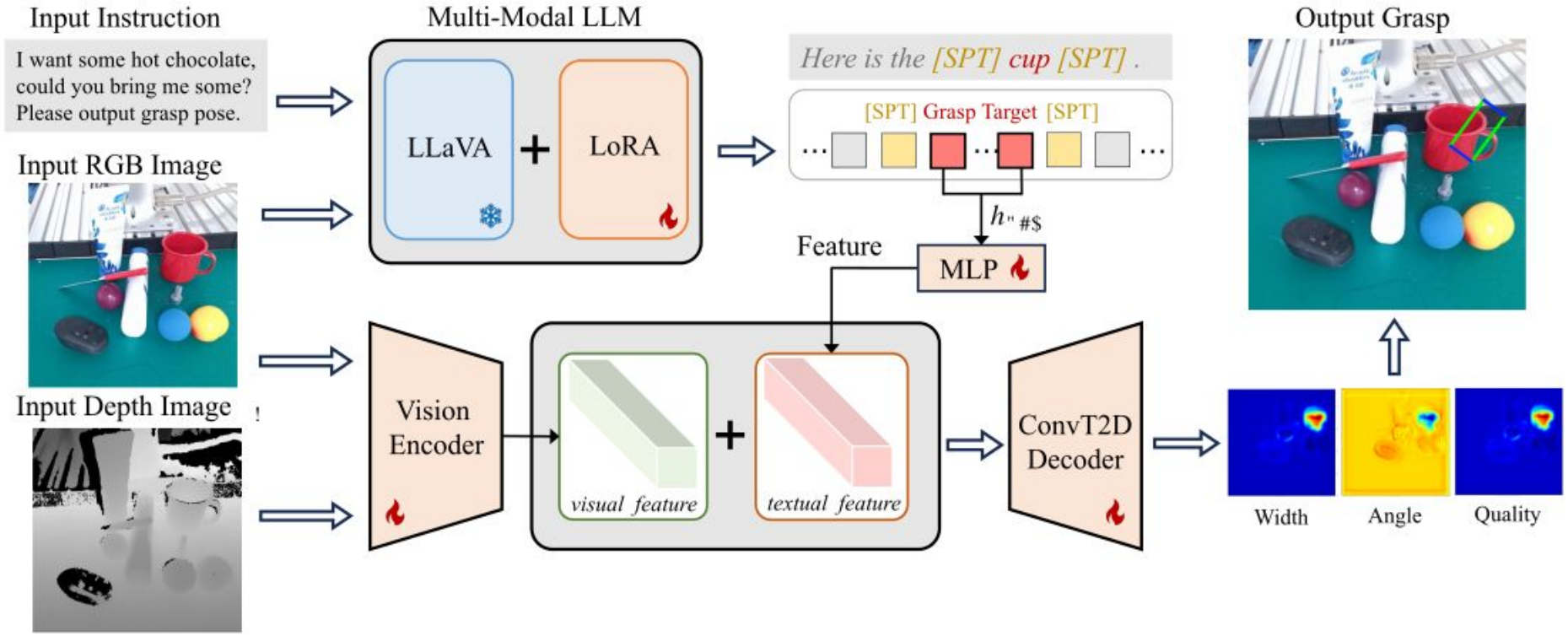
Taking argmax(Q), we get the best position where to grasp. Looking back to W and Φ, we can get the whole grasping pose.

Perception | Robotic Grasping

This is the result:

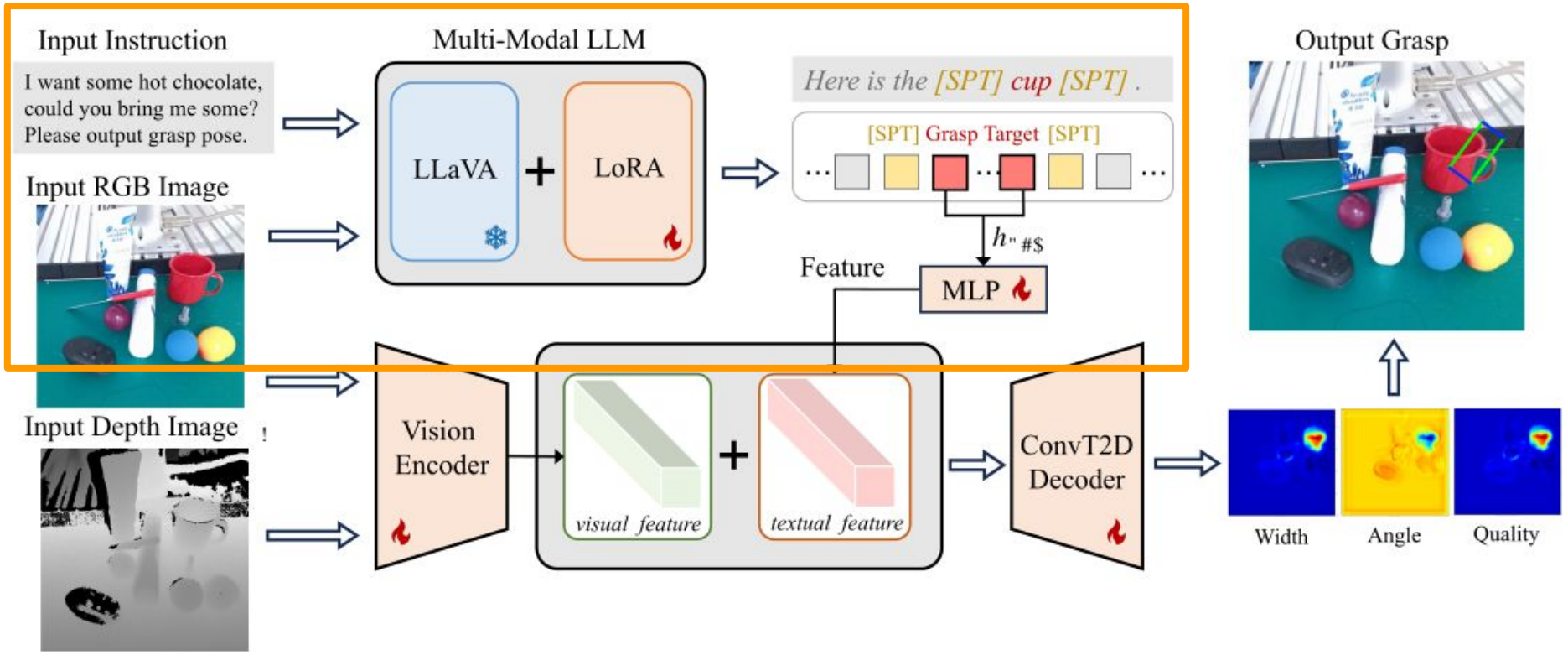


Perception | Next-level Robotic Grasping



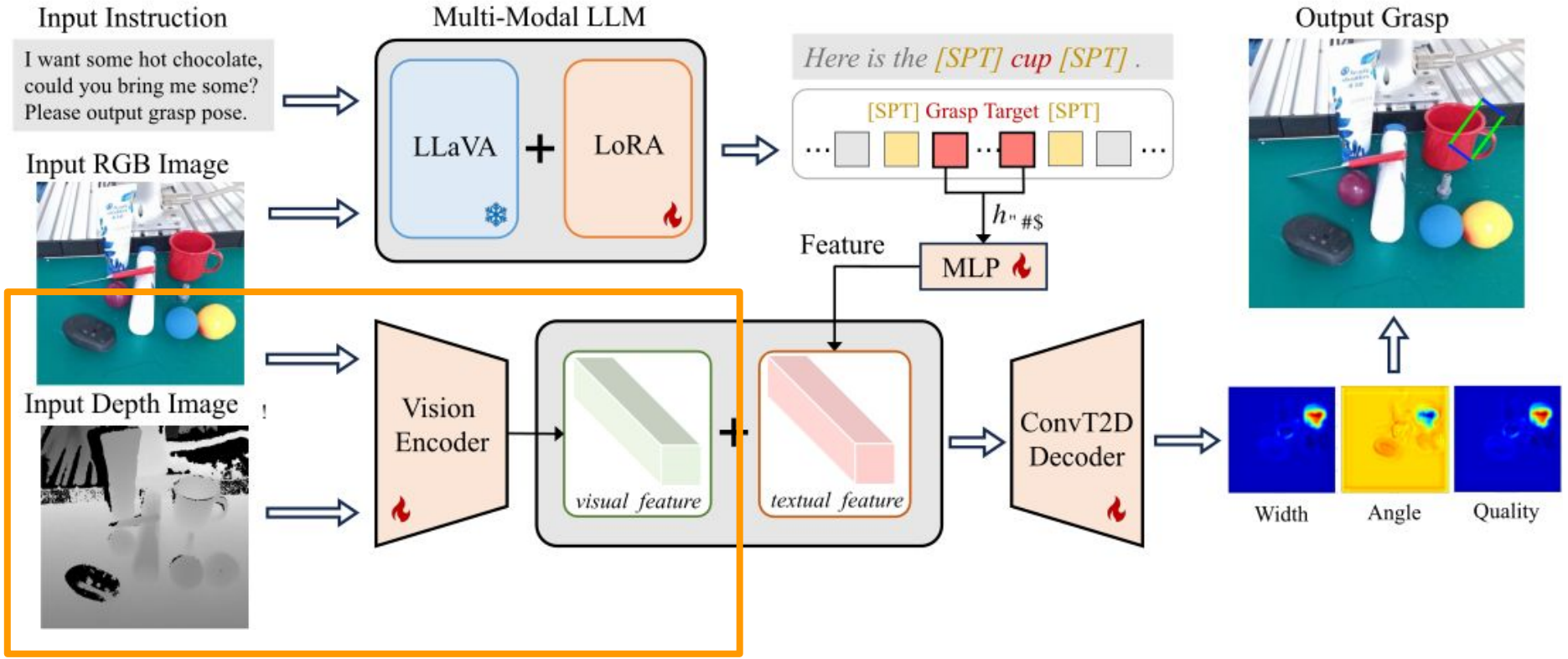
Mixing text and vision for (**implicit**) **targeted object grasping**.

Perception | Next-level Robotic Grasping



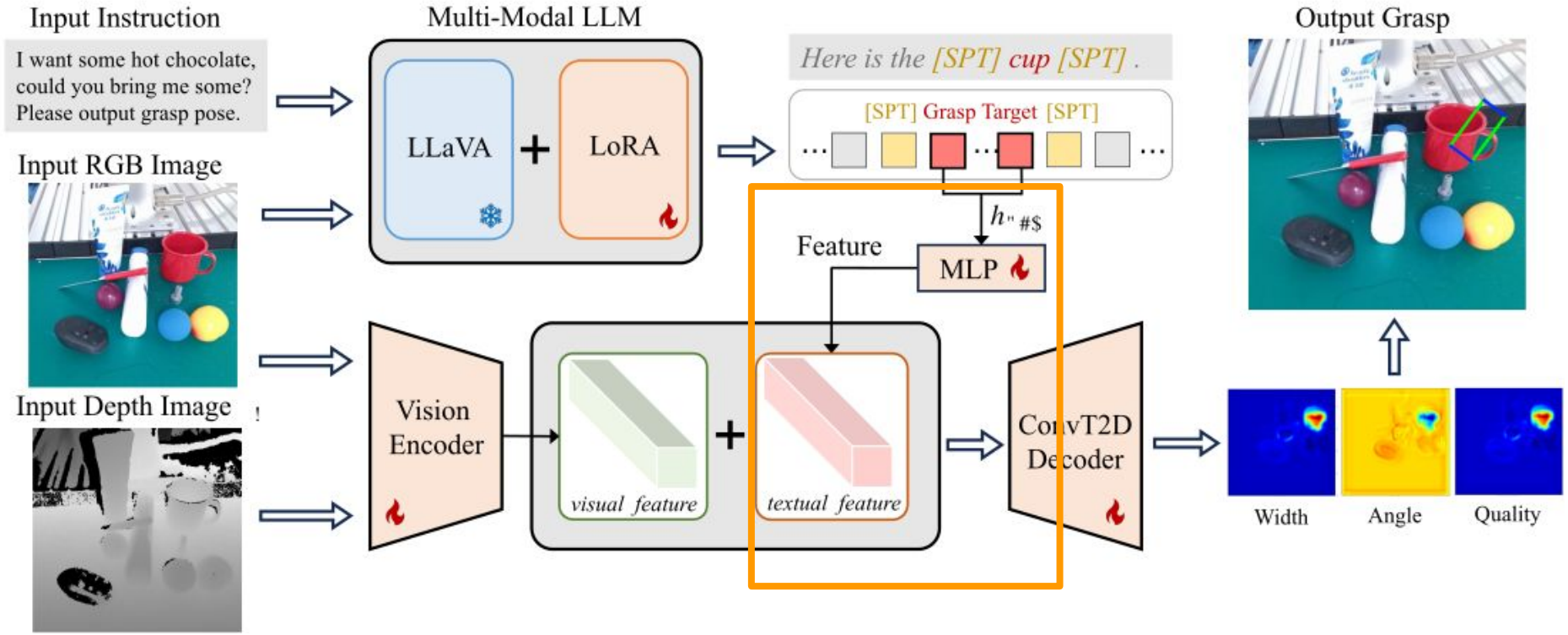
1. Fine-tune a LLaVa Vision-Language Model (VLM) to generate answers to the user query with special [SPT] tokens. [SPT] tokens isolate the target object.

Perception | Next-level Robotic Grasping



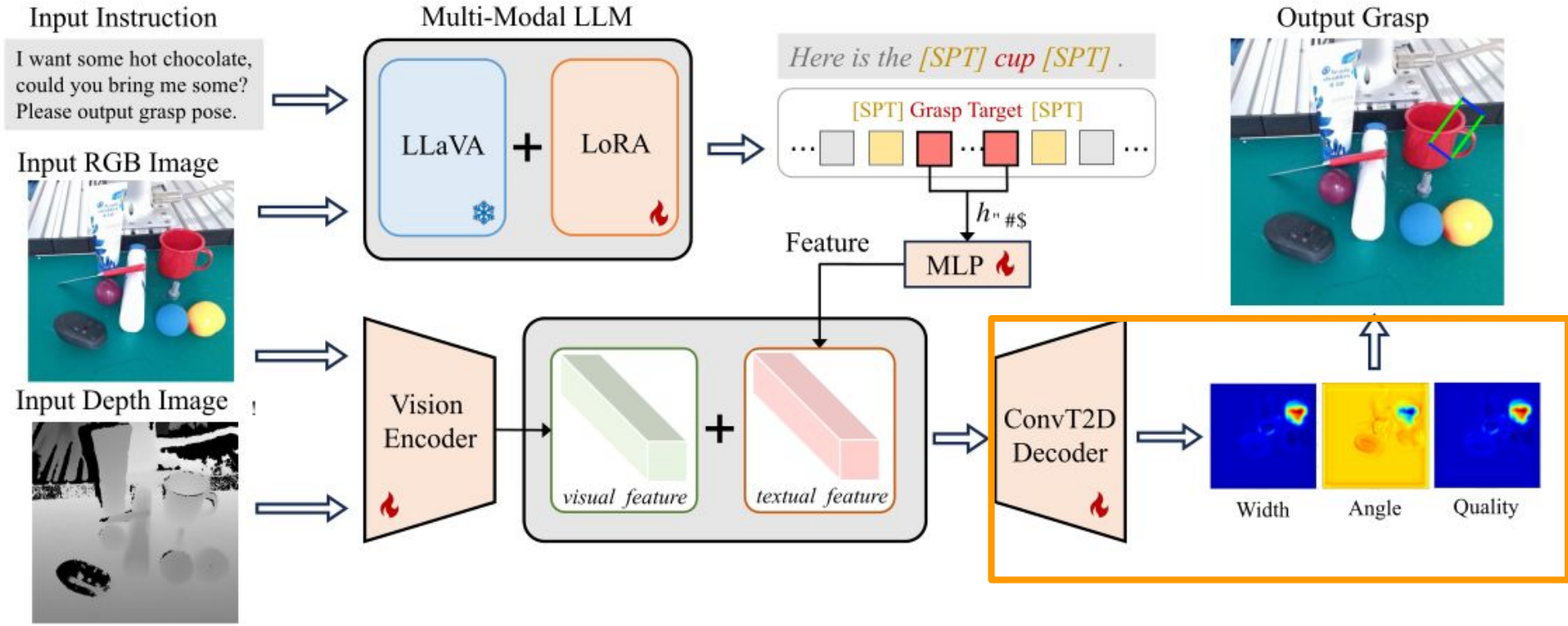
2. Train a Vision Transformer Encoder to extract visual information (embeddings) from the scene

Perception | Next-level Robotic Grasping



3. Map detected target into the sam embedding space of the visual features.

Perception | Next-level Robotic Grasping



4. Decode embedding to generate target-oriented heatmaps.

Perception | Next-level Robotic Grasping

This is the result:

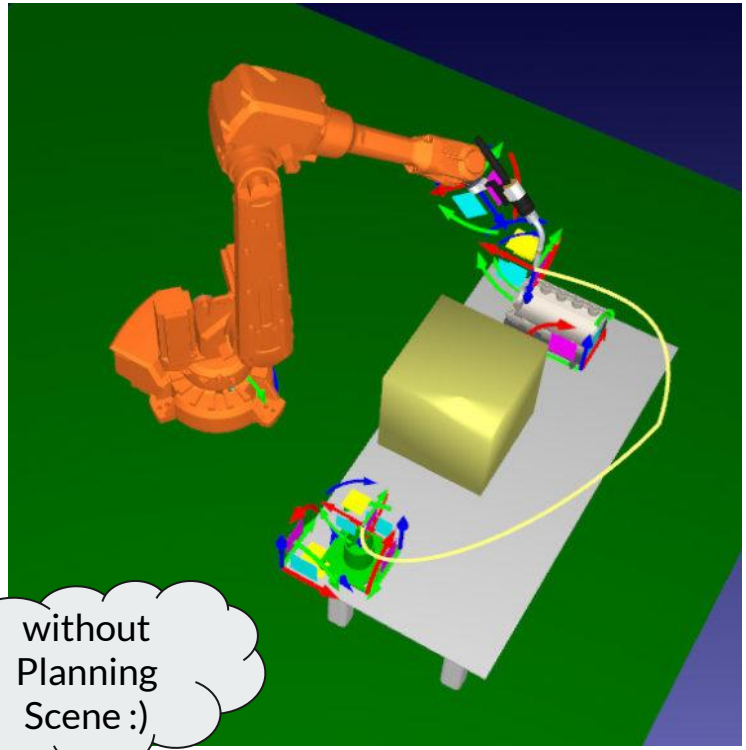
Reasoning Grasping via Multimodal Large Language Model

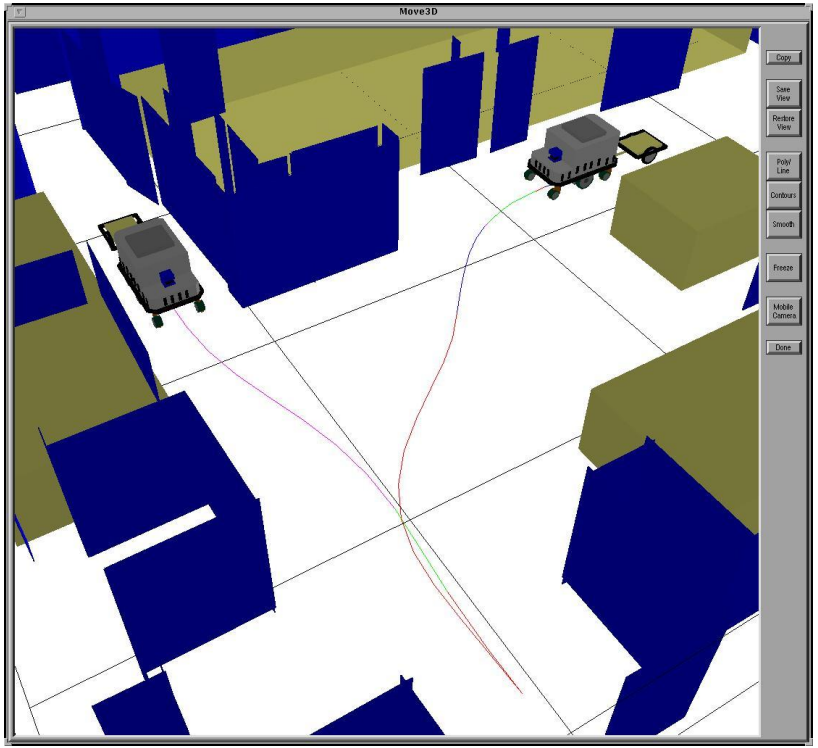
Author Names Omitted for Anonymous Review – Submission Number 439

Planning

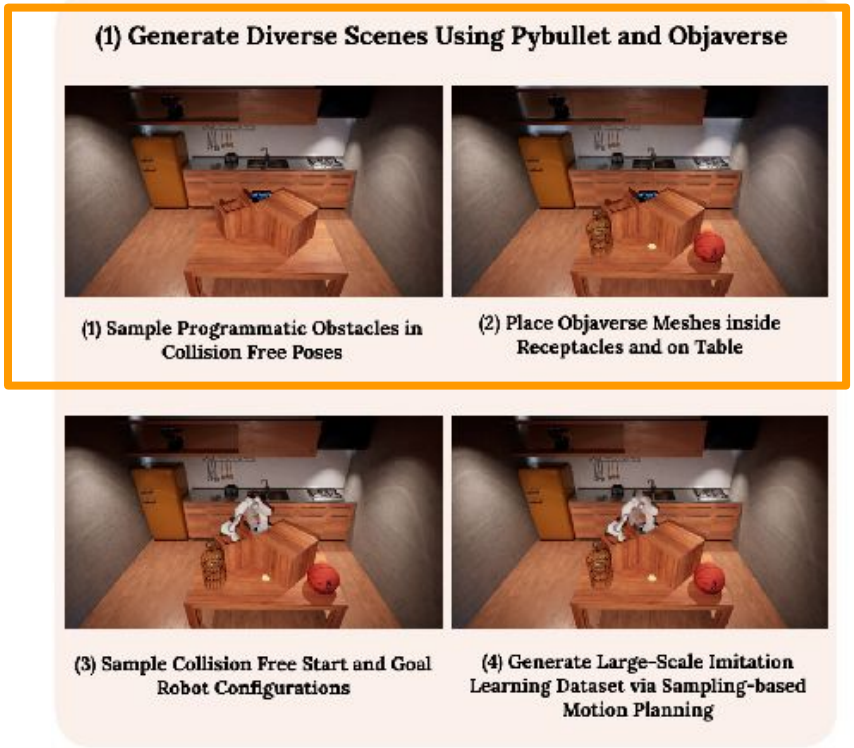
The task of generate suitable trajectories for the robot.

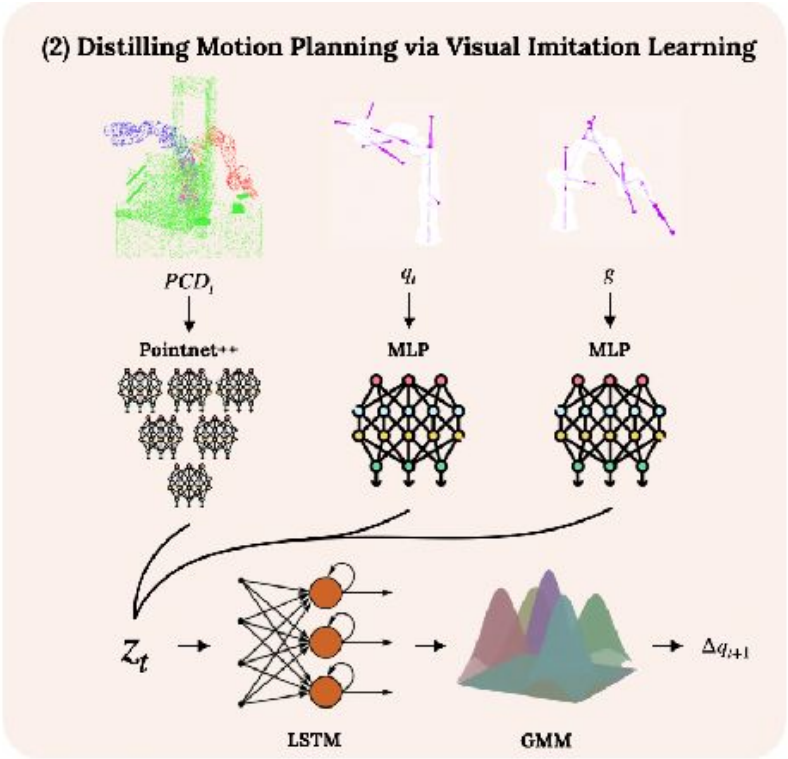
Suitable means feasible (no collisions, within kinematics limits) and successful (the robot can reach the goal).





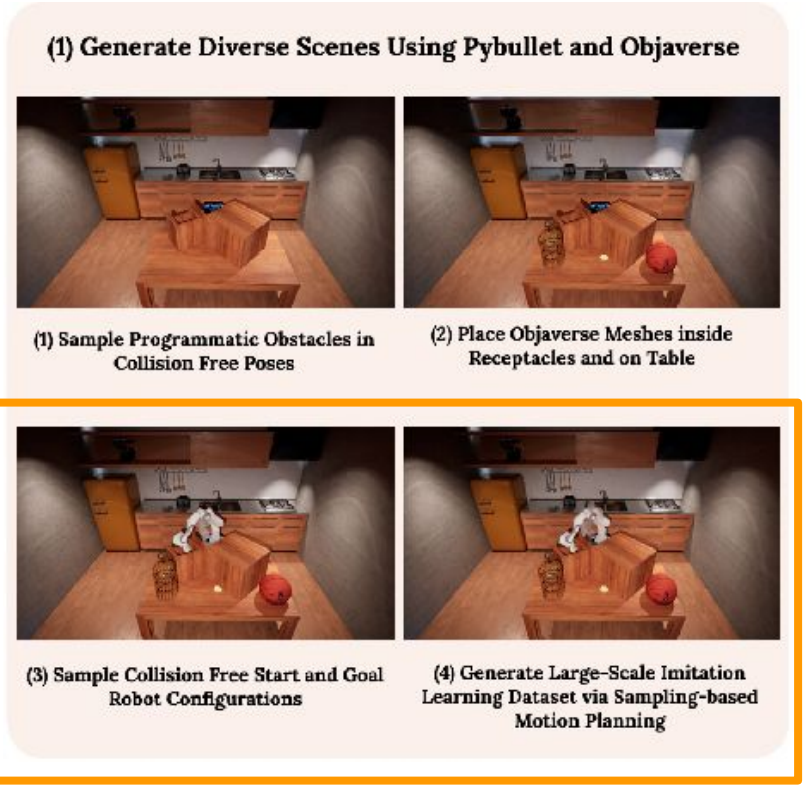
Planning | Neural Motion Planning

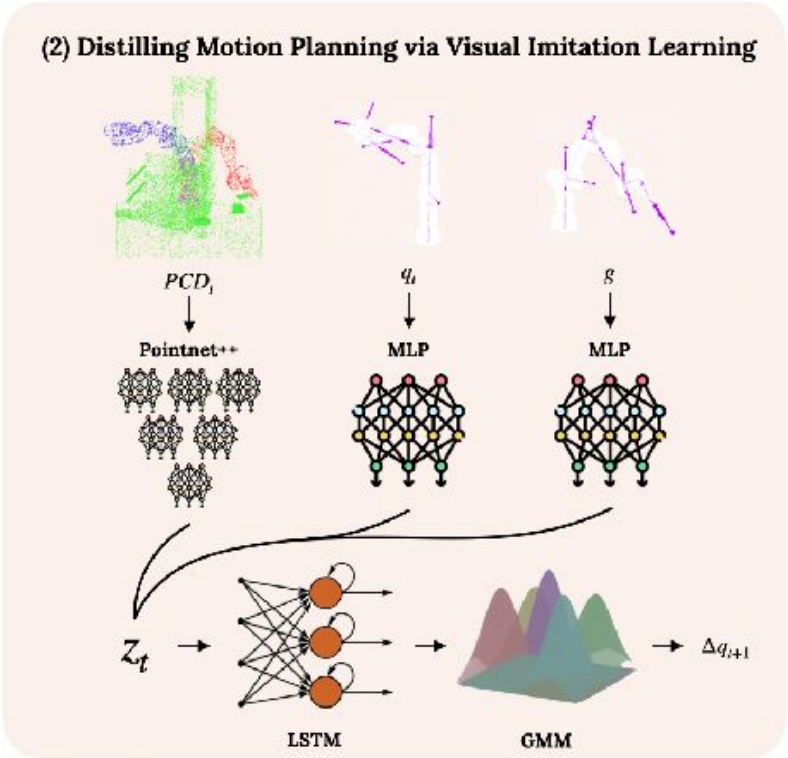




1. Generate synthetic scenes using a simulator and public datasets

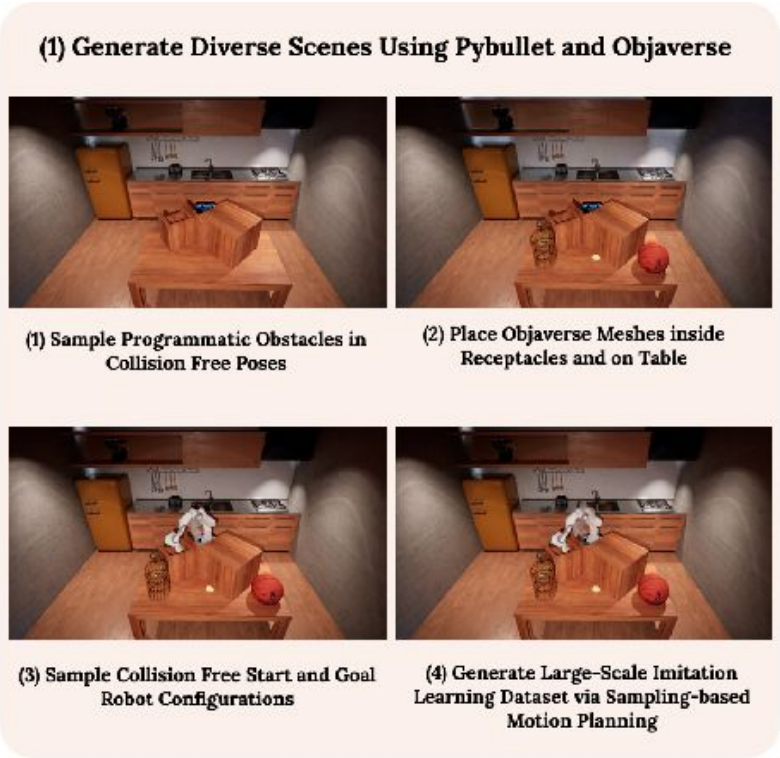
Planning | Neural Motion Planning

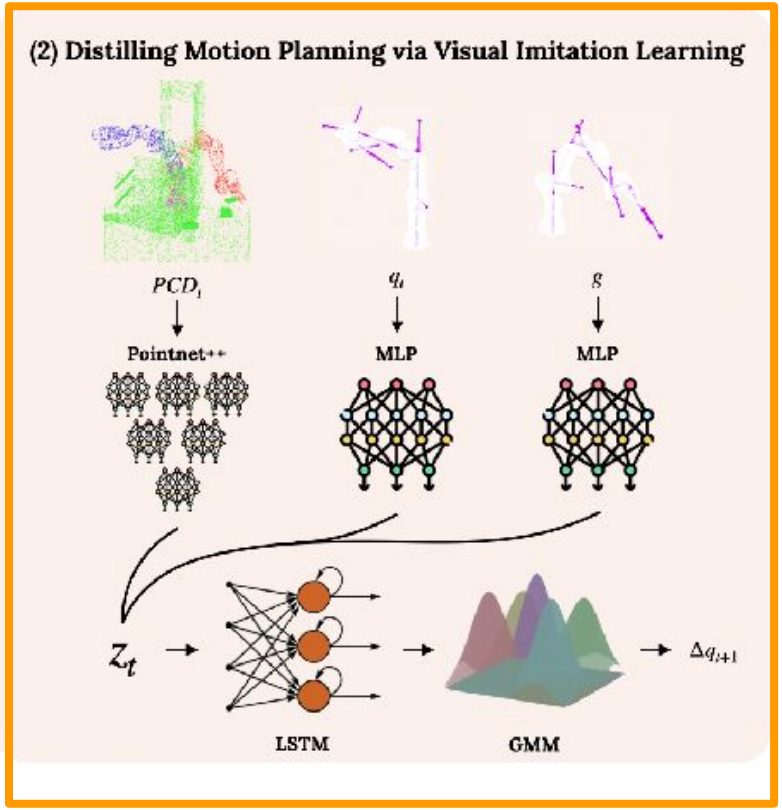




2. Sampling-based motion planners to generate training data (trajectories). Note that in simulation it is easier to generate high-quality data (perfect collision checking, offline optimization, ecc…)

Planning | Neural Motion Planning





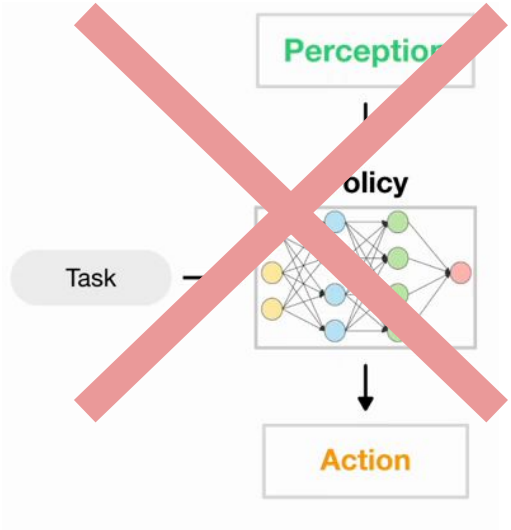
3. Train a deep neural network to predict trajectories on synthetic data. Inputs are the point cloud, the current robot joint state and the robot target pose. Output is the next joint state.

Planning | Neural Motion Planning

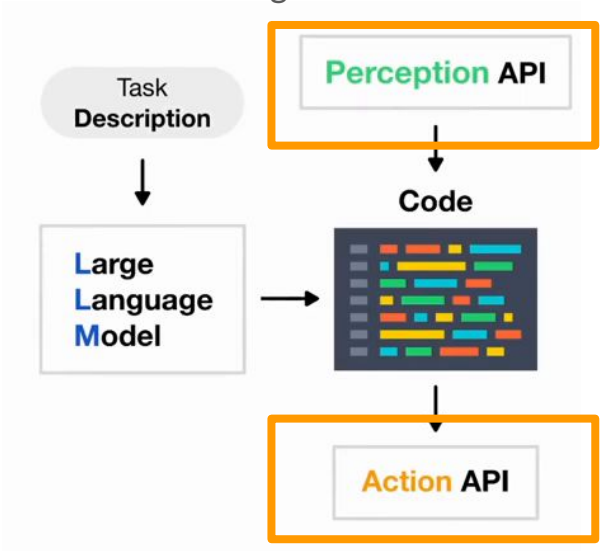
This is the result:

Planning | Task Planning with LLM

Before: task-specific models



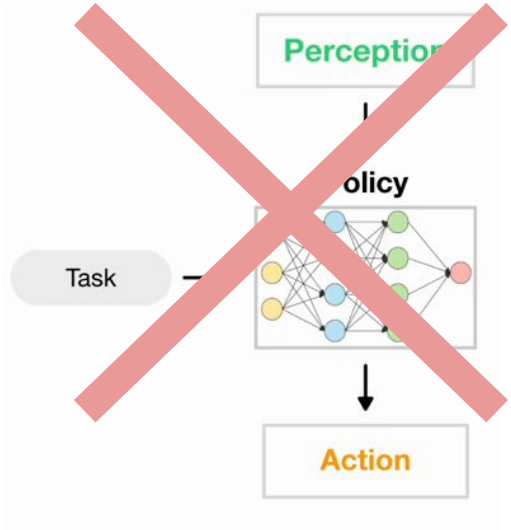
Now: task-agnostic models



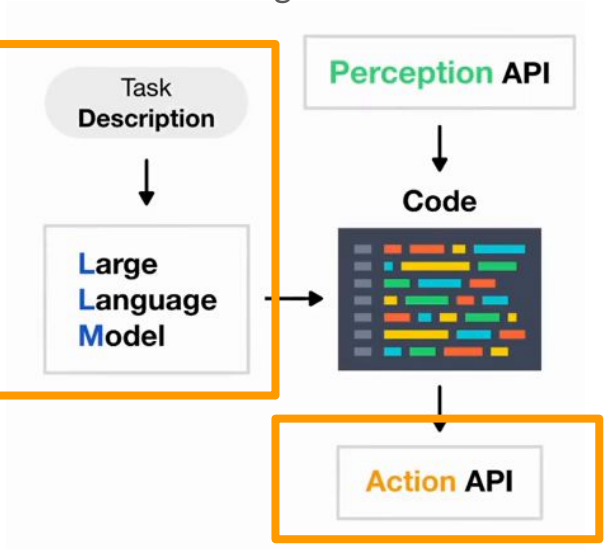
1. Define some atomic operations implemented by some APIs, for example `acquireImage()`, `moveTo()`, ecc...

Planning | Task Planning with LLM

Before: task-specific models



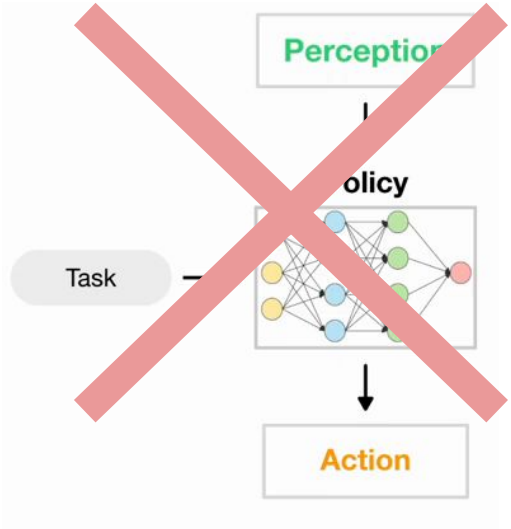
Now: task-agnostic models



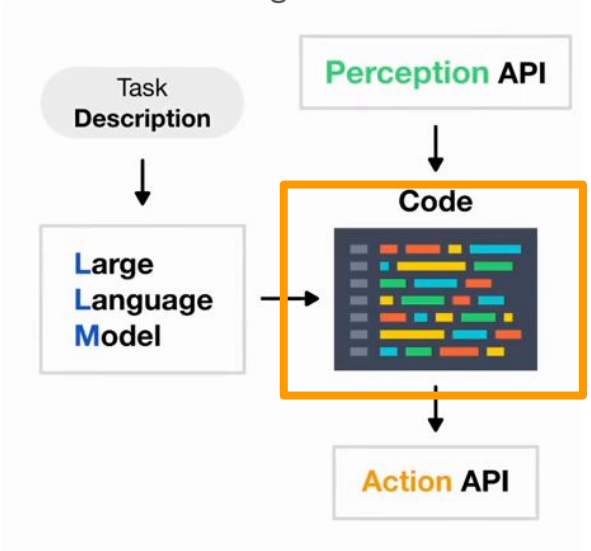
2. Fine-tune a LLM to translate a task description into code, using the given API

Planning | Task Planning with LLM

Before: task-specific models



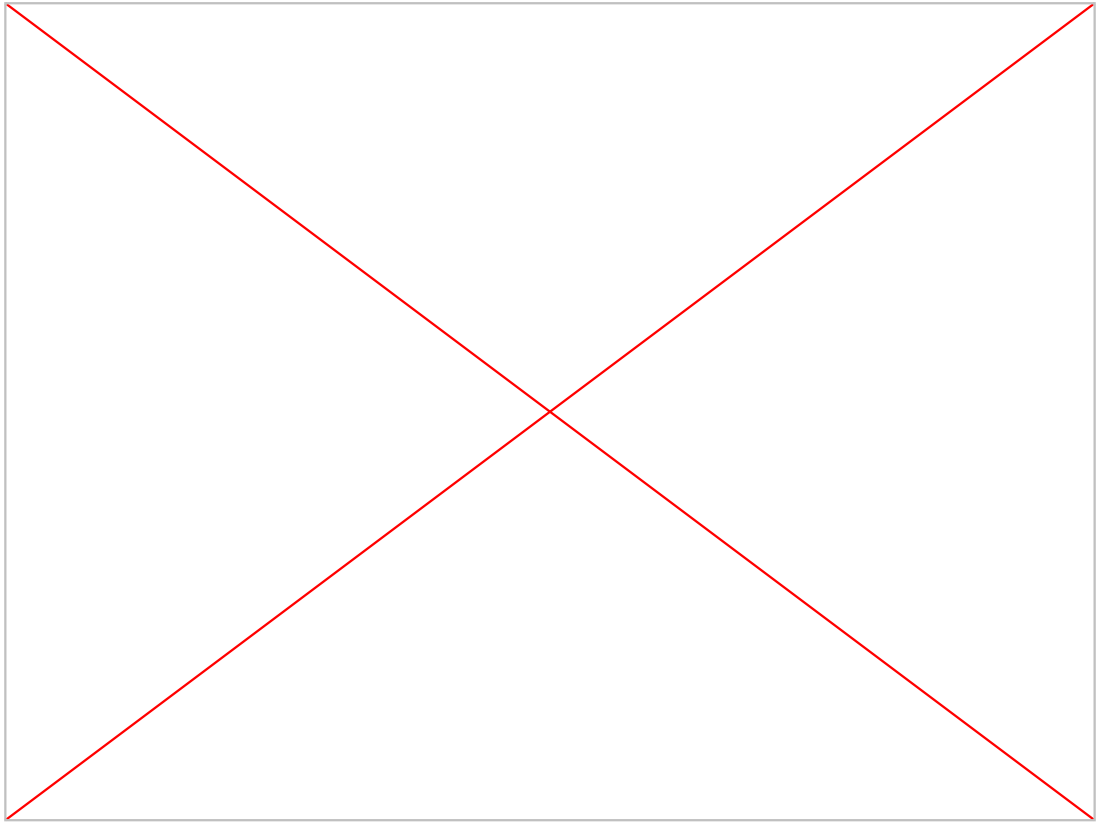
Now: task-agnostic models



3. Execute the code to achieve the task

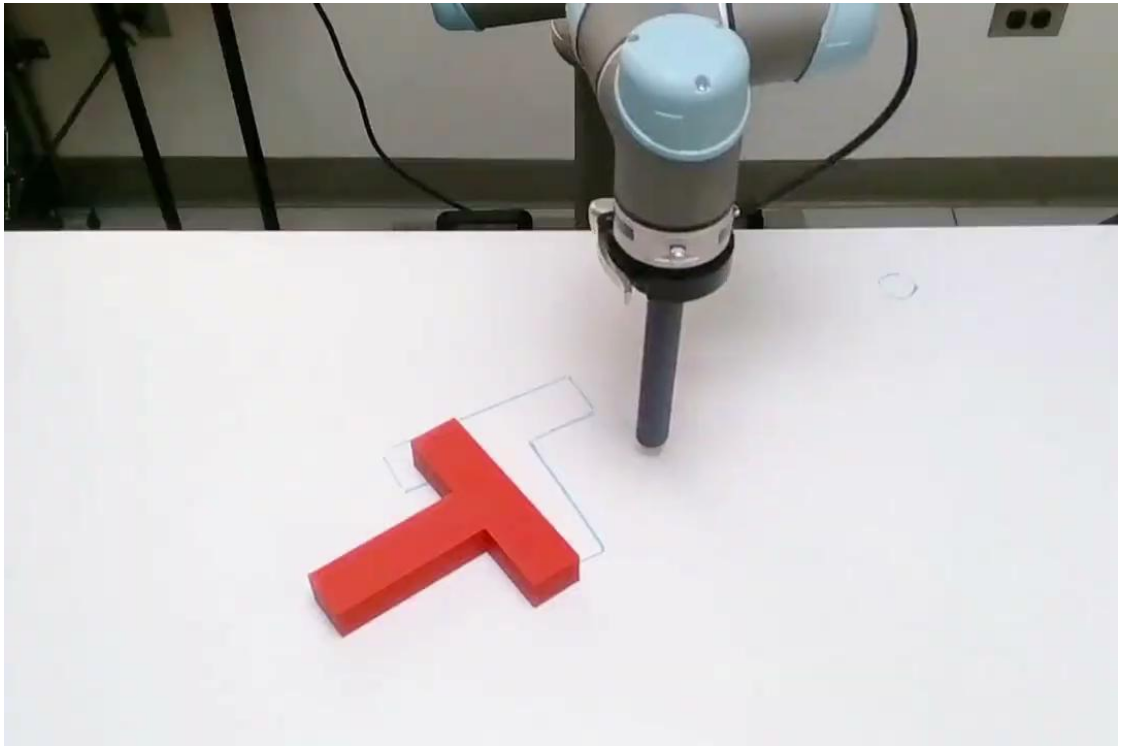
Planning | Task Planning with LLM

This is the result:



Planning | Planning with Diffusion Models

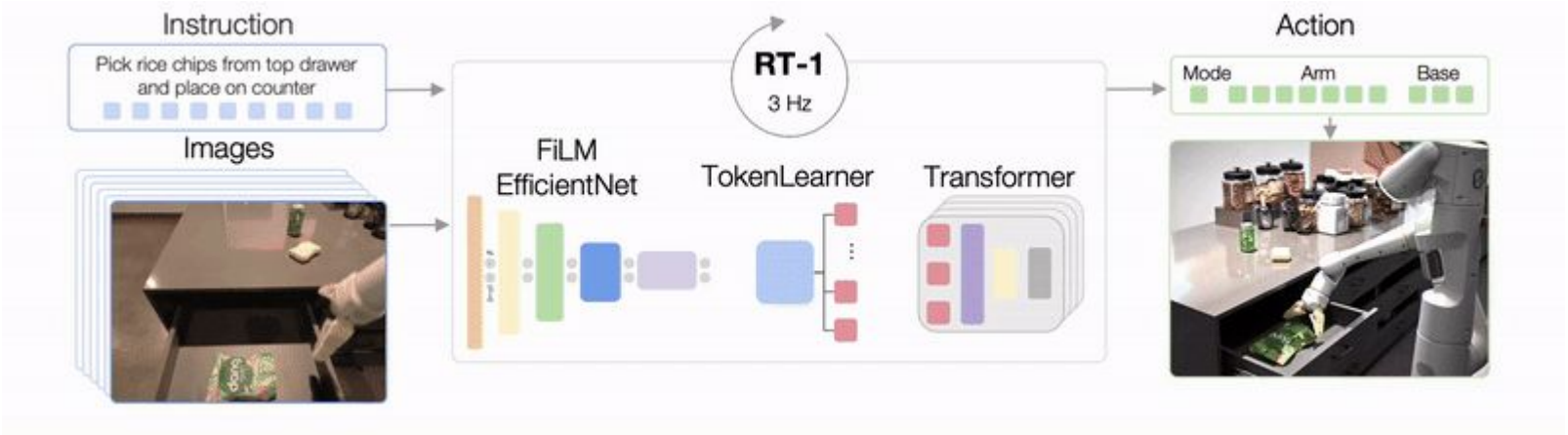
We do not cover Diffusion Models, but they are powerful tools for robotics:



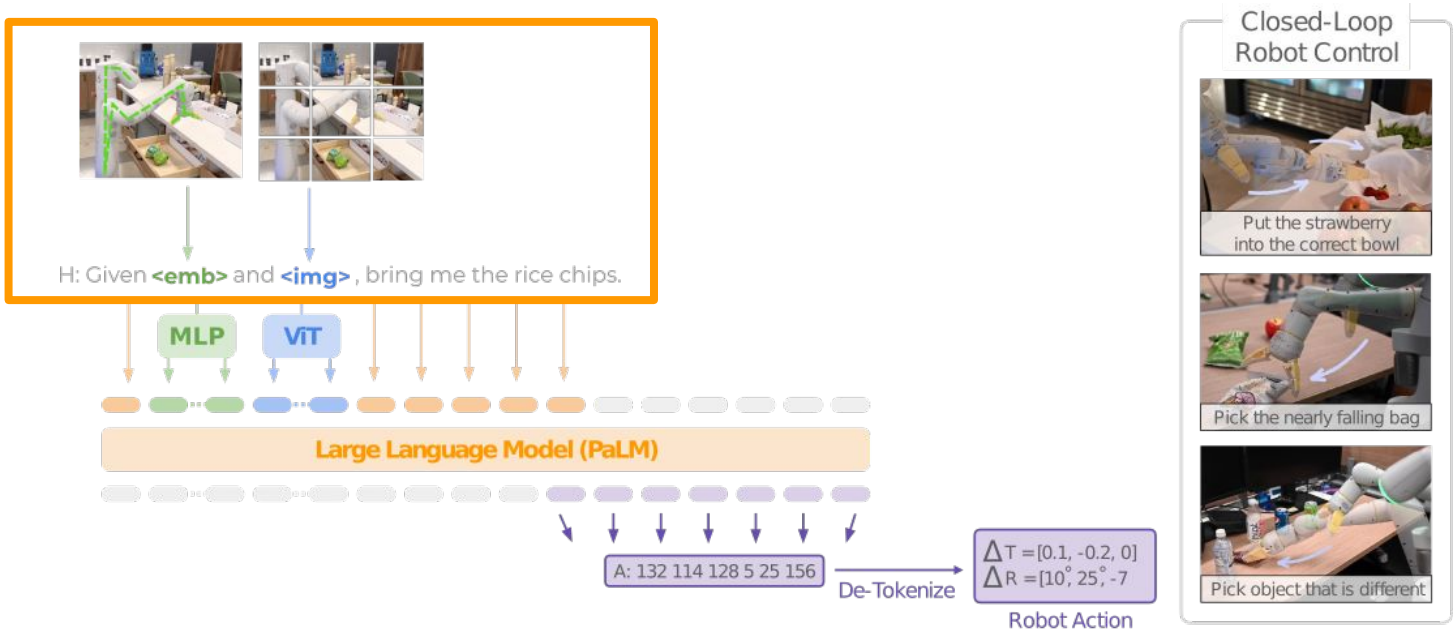
Sense-Plan-Act AiO

Input raw sensor readings and task description and output robot control signals directly.

It is an end-to-end “real-time” robot controller.

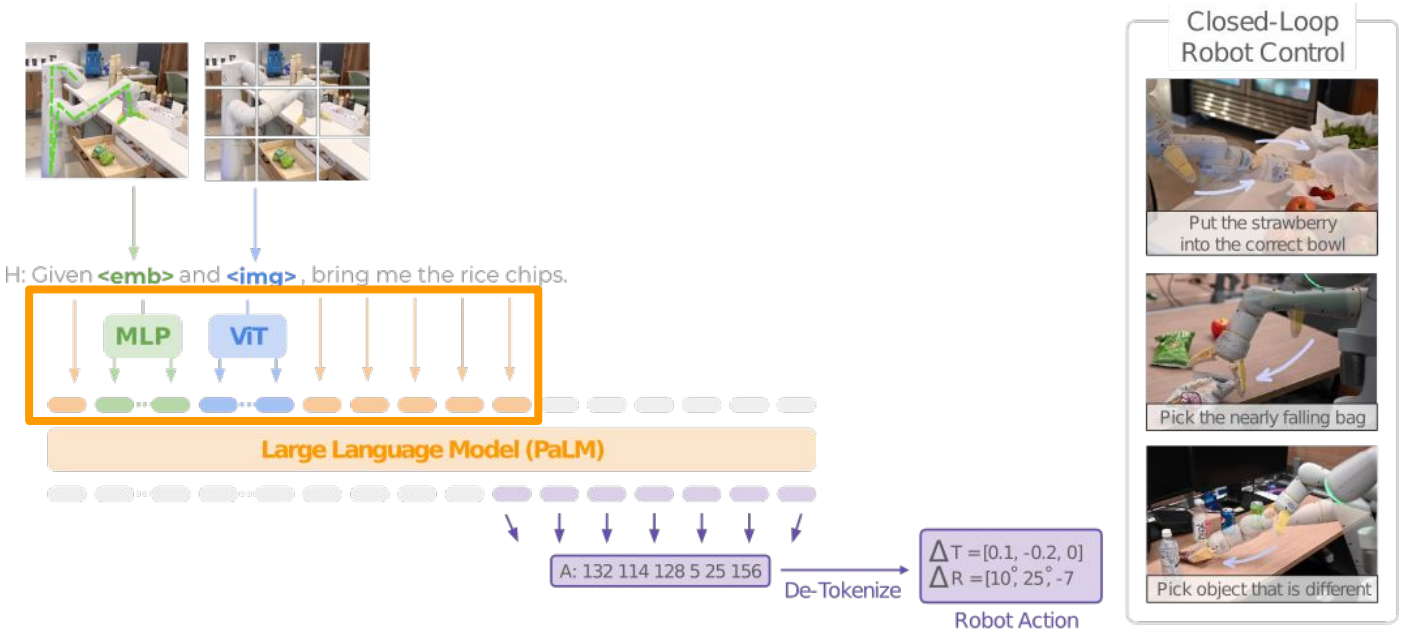


Sense-Plan-Act AiO | Robotics Transformer



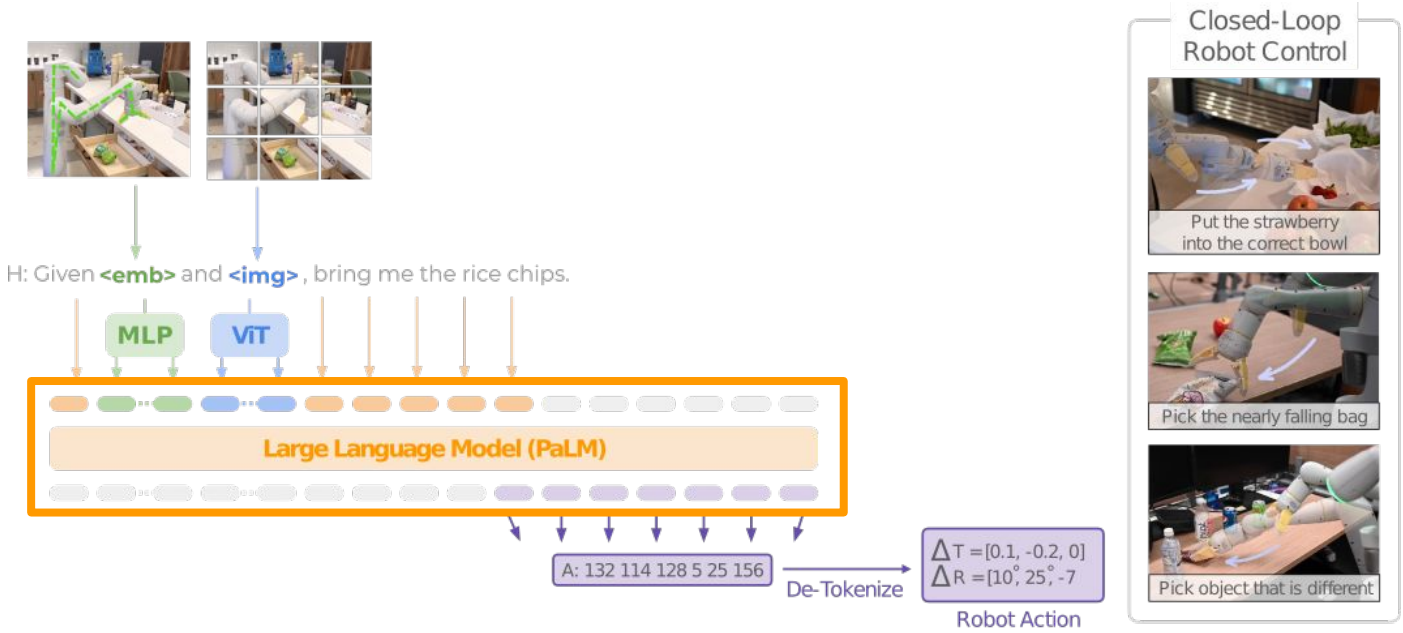
1. Prepare a **natural language prompt with "special placeholder"** for sensor data (images, robot proprioception, ecc…) and the task description.

Sense-Plan-Act AiO | Robotics Transformer



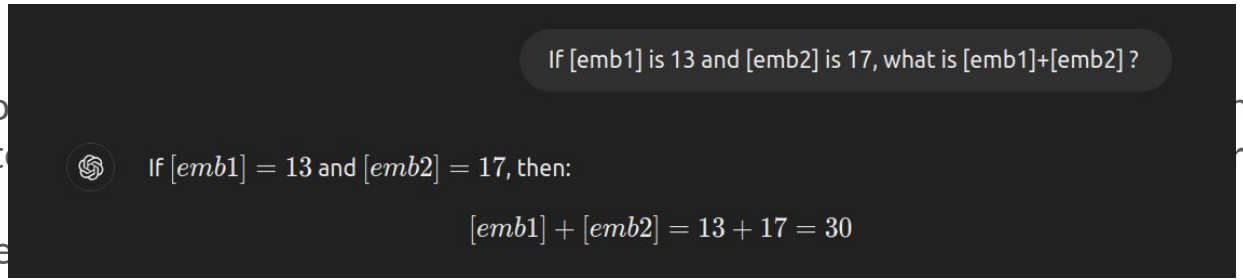
2. Tokenize your input. Use **specific tokenizer for each mode.** For example Vision Transformer for images.

Sense-Plan-Act AiO | Robotics Transformer

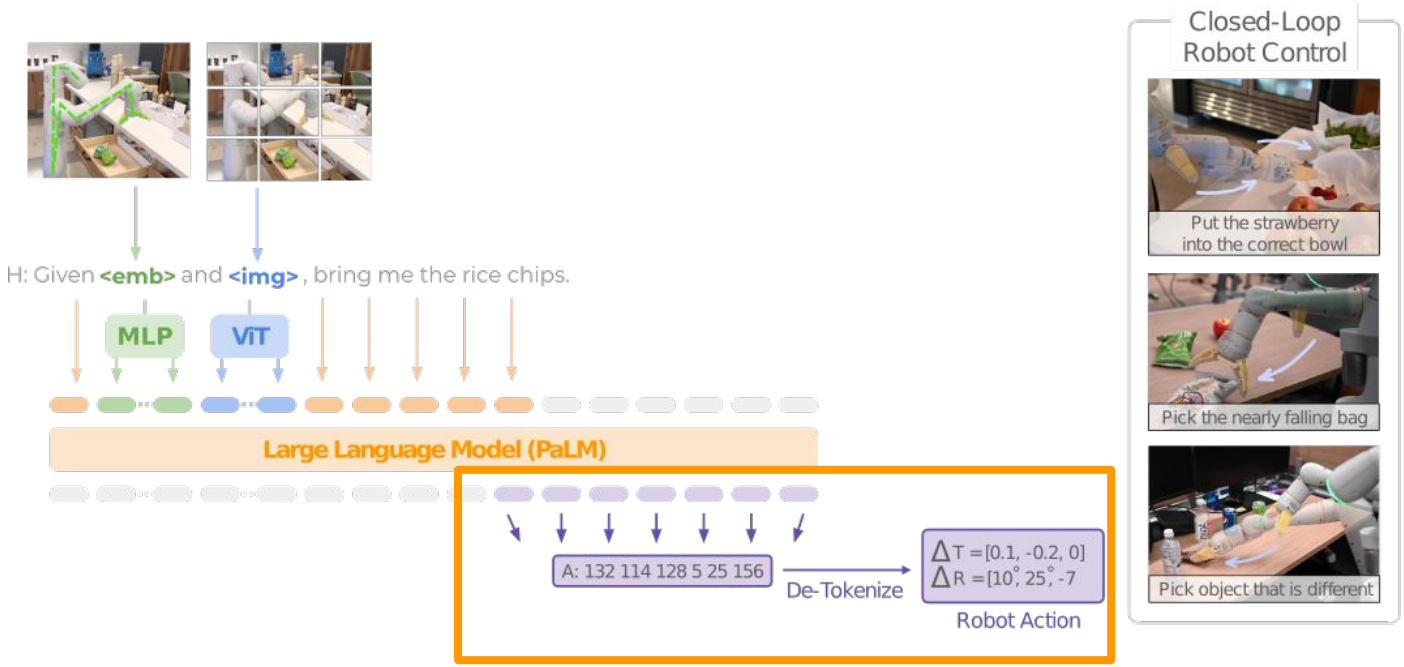


3. Use LLM to … hat the input is a sequence of t… repeat.

LLM have bee…

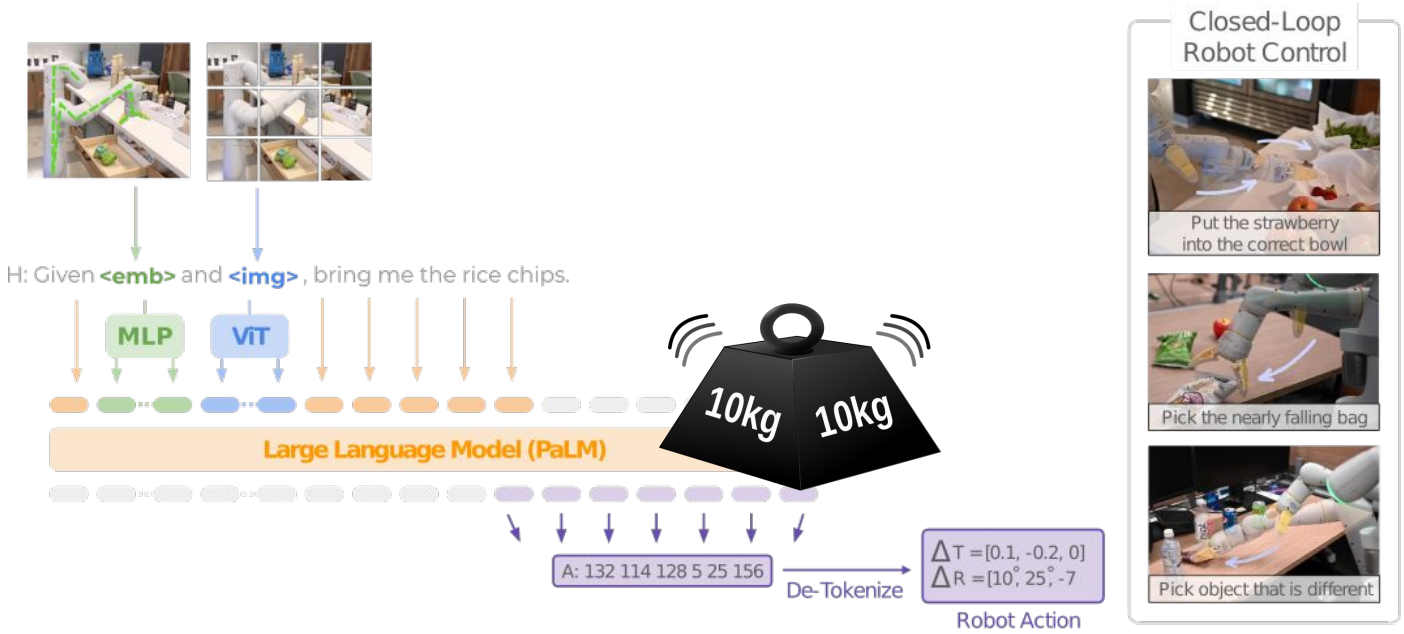


Sense-Plan-Act AiO | Robotics Transformer



4. Fine-tune the LLM to **output tokens that represents action** in the robot's space. De-tokenize actions.

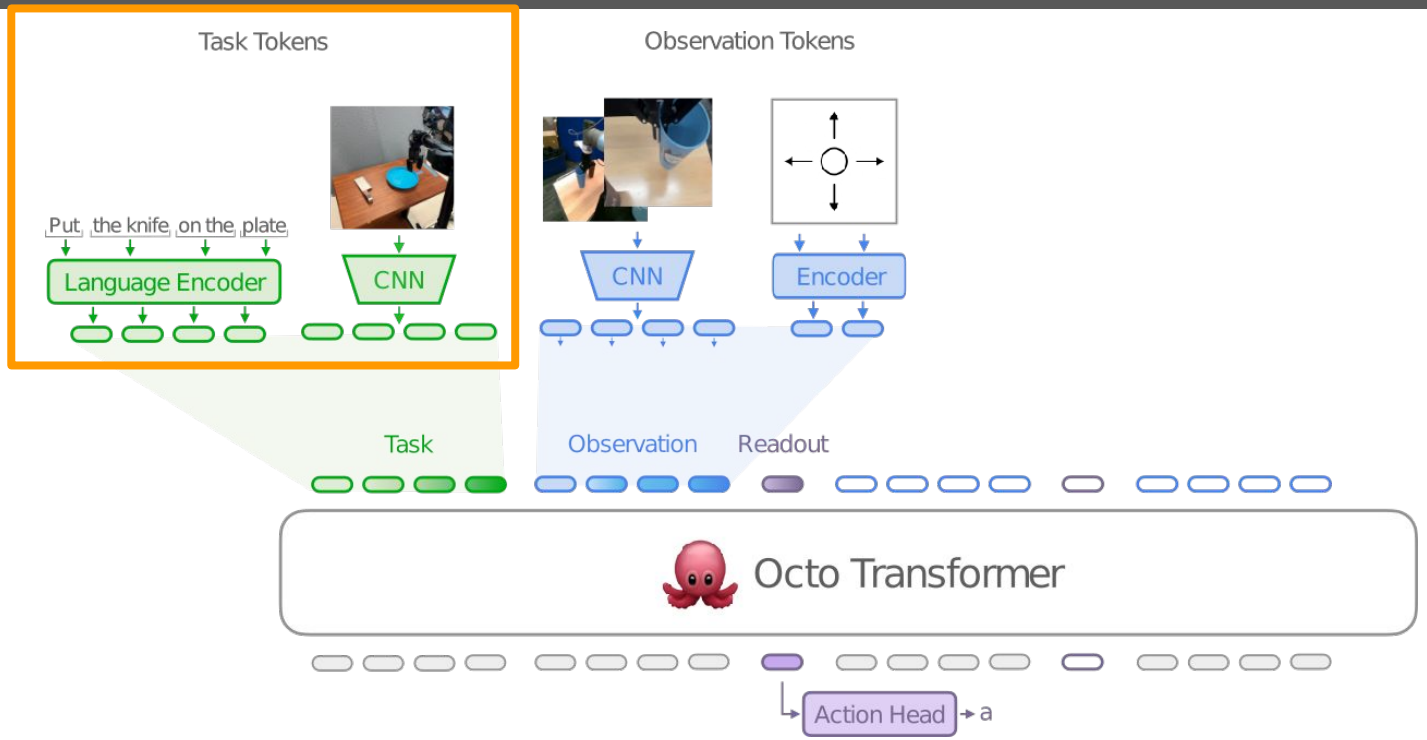
Sense-Plan-Act AiO | Robotics Transformer



PROBLEM: LLM are powerful, but very heavy (PaLM ~ 540B parameters).

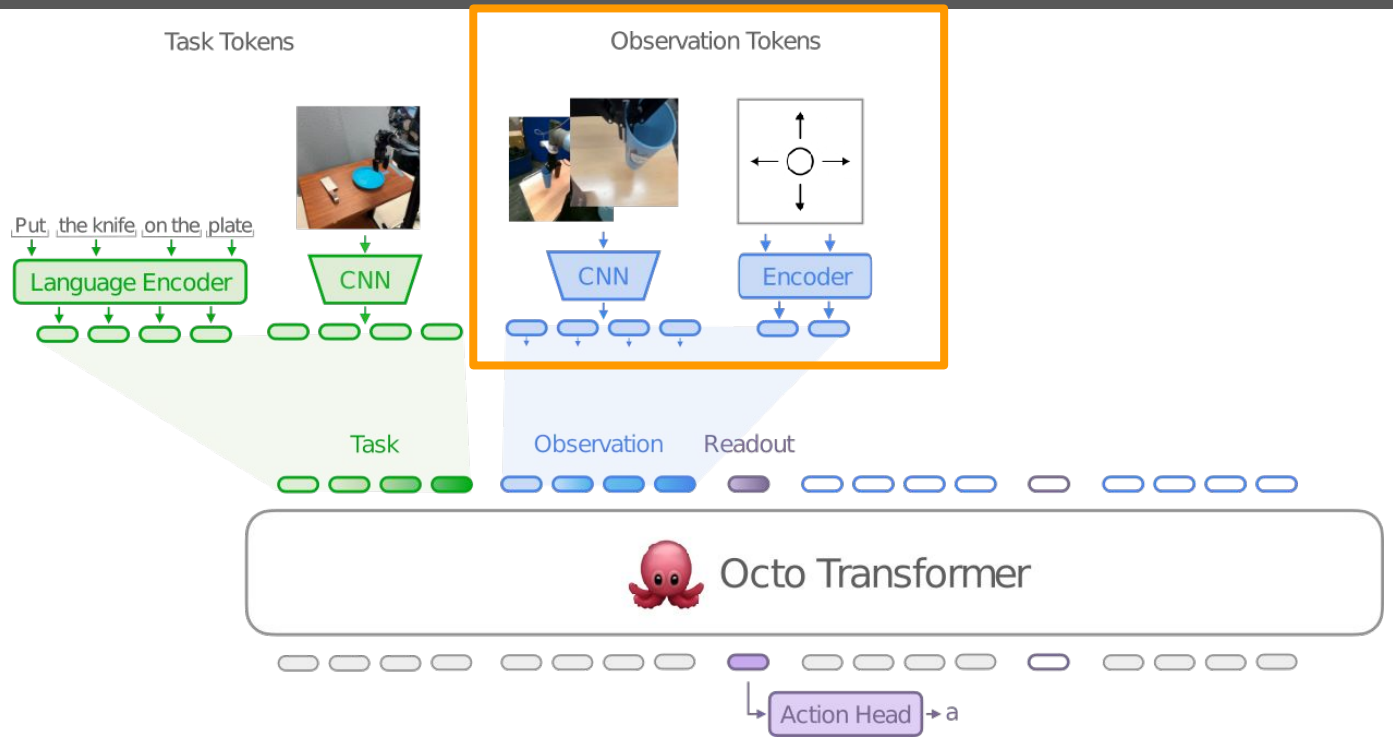
IDEA: PaLM is a generic agent, not optimize in particular for robotics…

Sense-Plan-Act AiO | Light Robotics Transformer



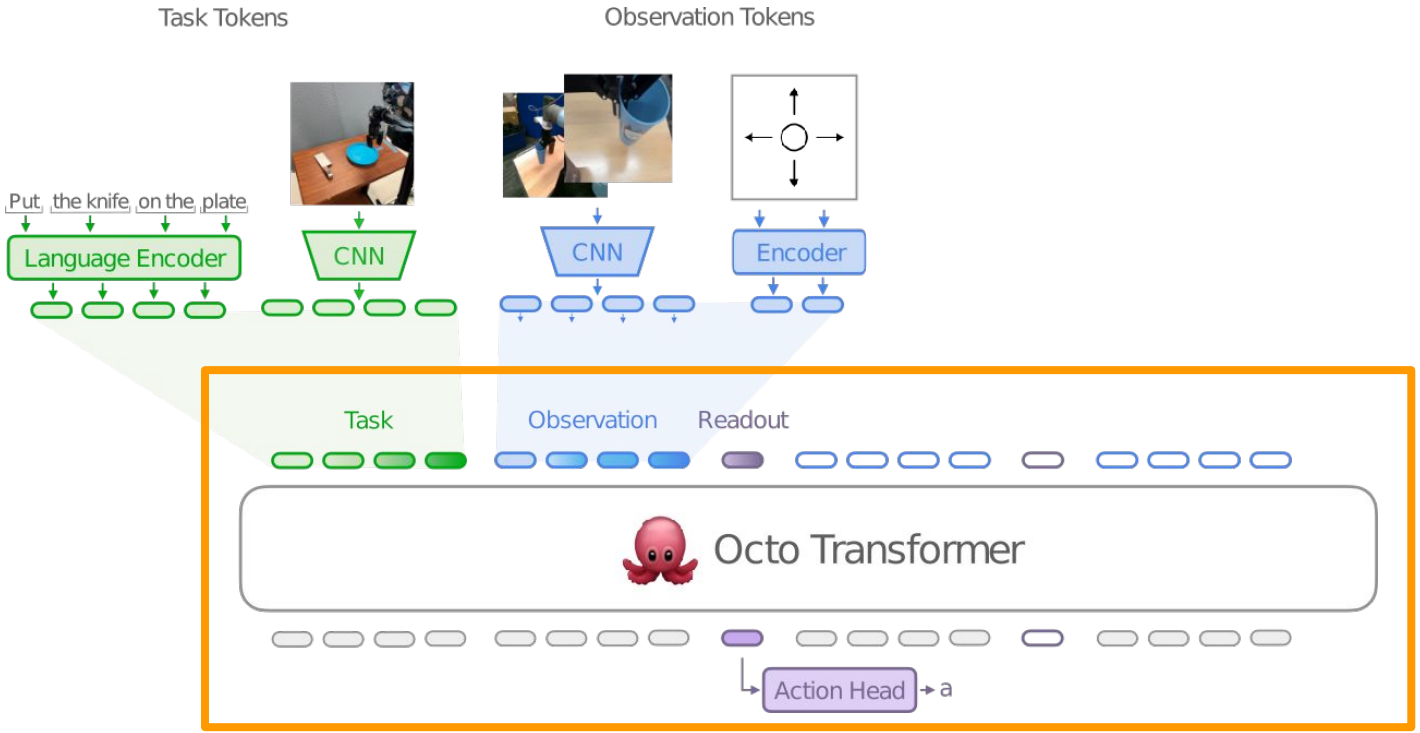
1. **Define the task** using language and/or images. **Map task description into tokens** using **light encoders** for each modality.

Sense-Plan-Act AiO | Light Robotics Transformer



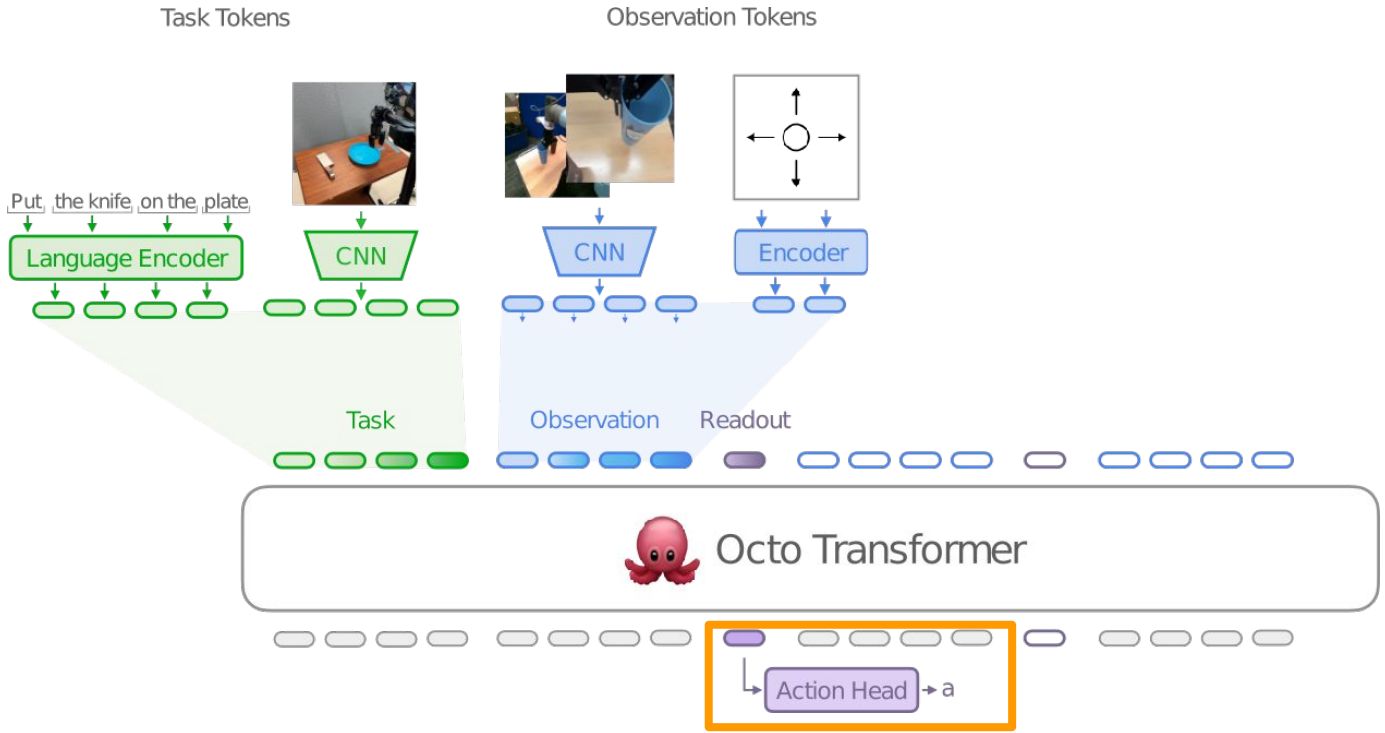
2. **Map the observations into tokens** using light encoders.

Sense-Plan-Act AiO | Light Robotics Transformer



3. **Train an (encoder) transformer** to predict action embeddings.

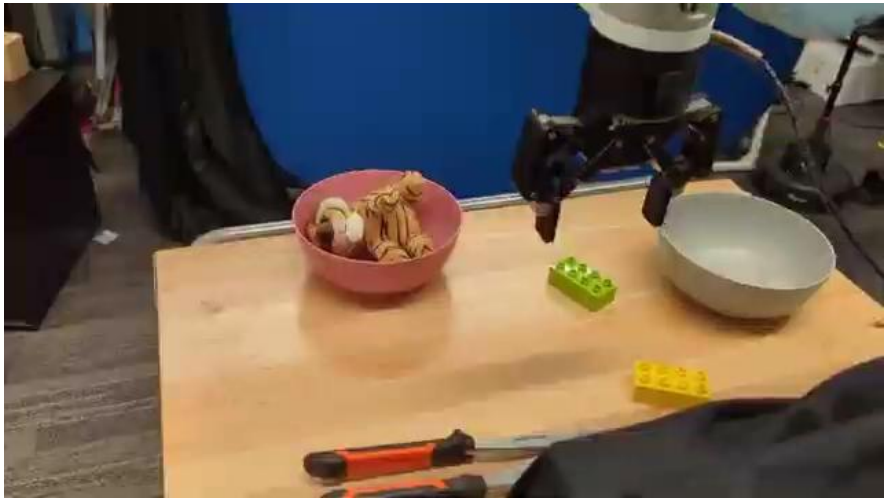
Sense-Plan-Act AiO | Light Robotics Transformer



4. **Decode the action embeddings** with a light action head.

Sense-Plan-Act AiO | Light Robot Transformer

This is the result:



Compared to Large Robotic Transformer:

- More efficient (~200M params vs ~550B)
- Modular
- Less Reasoning and Zero-Shot capabilities (needs fine-tuning)

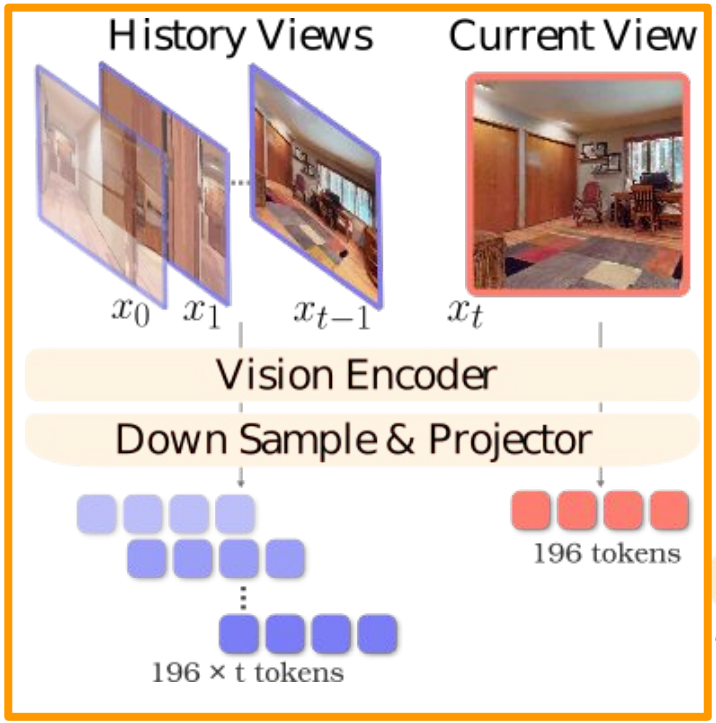
Sense-Plan-Act AiO | Navigation with VLA

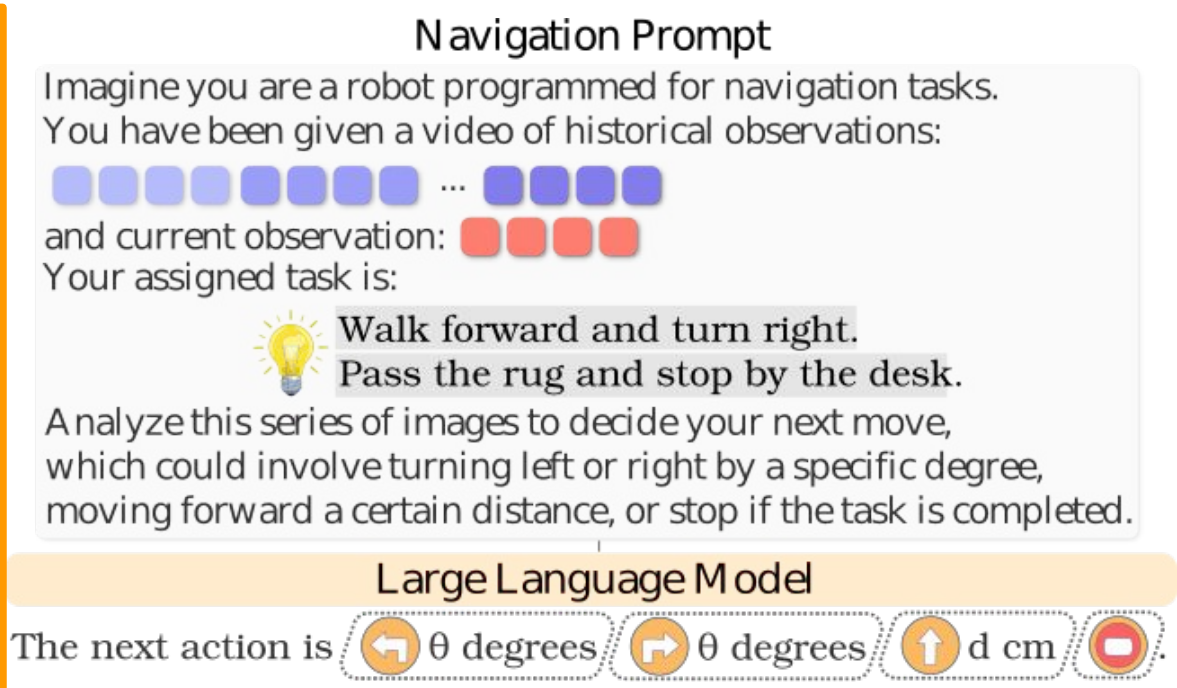
Use Vision Language Action models to generate navigation subgoals, given instructions

Navigation subgoals are the actions to be executed. Note that Robotic Transformer is a VLA as well.

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

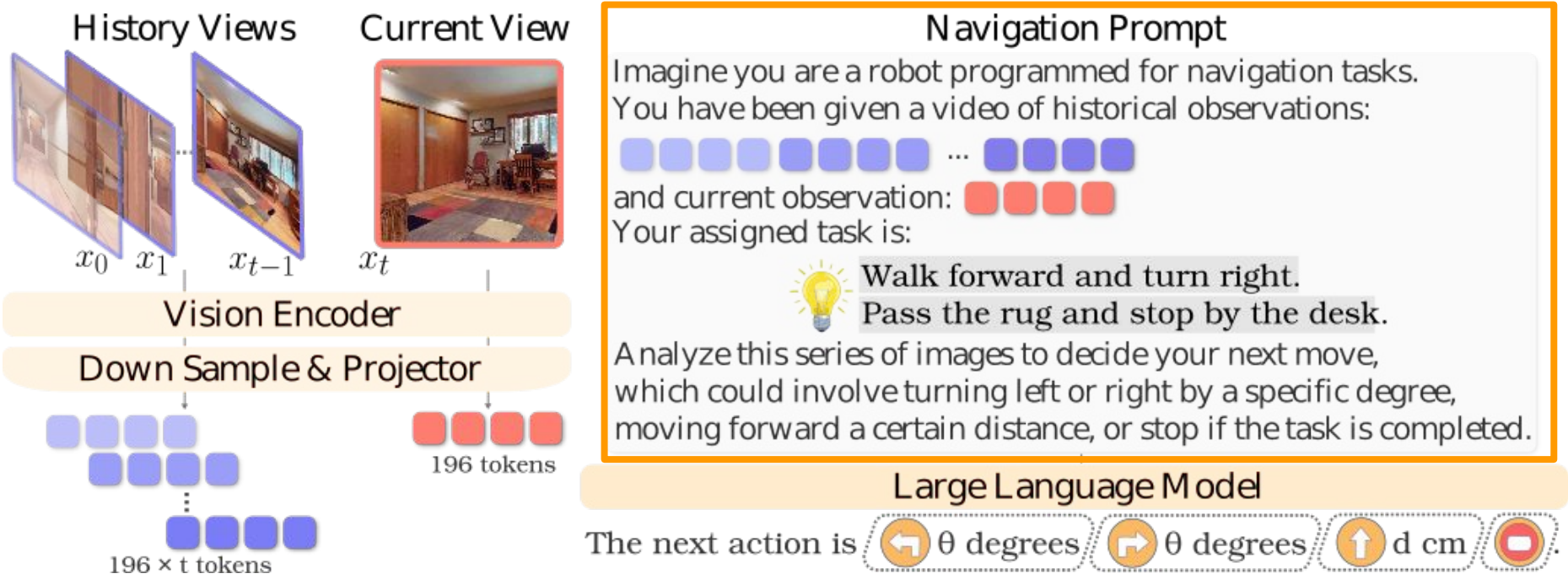
Sense-Plan-Act AiO | Navigation with VLA





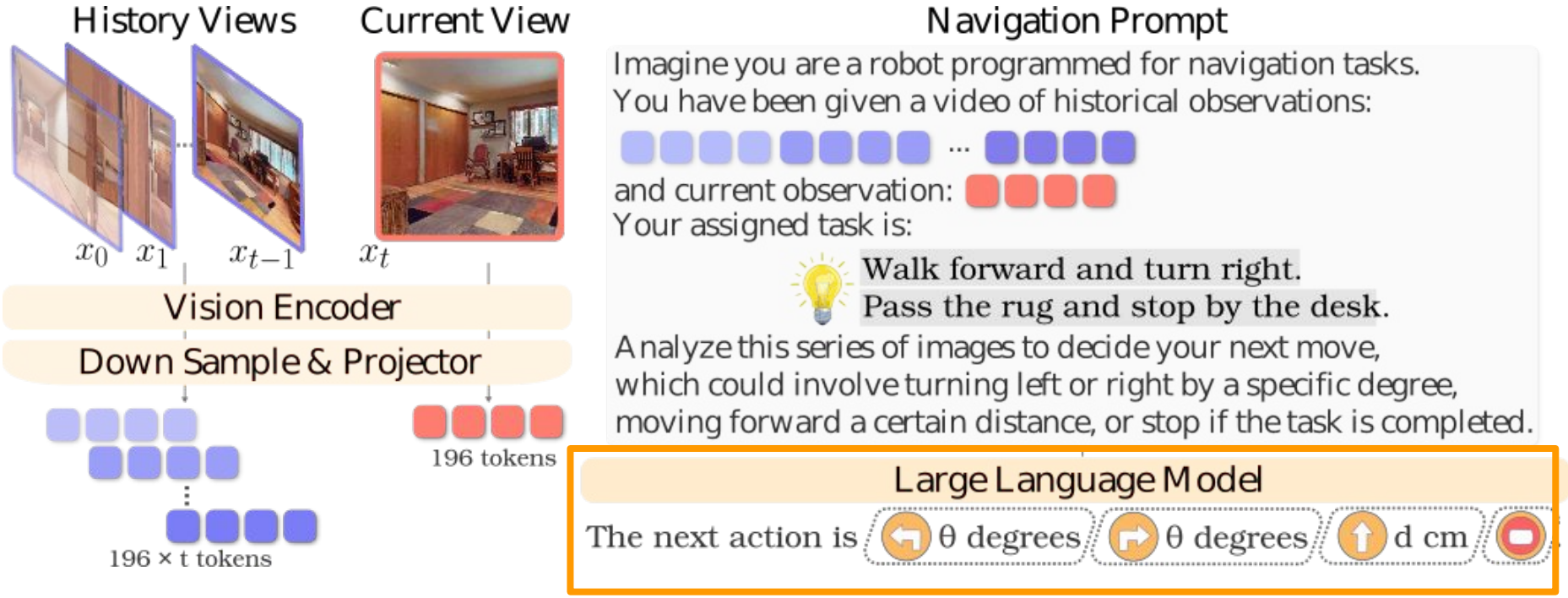
1. A Vision Transformer Encoder is used to **map images (previous views and current view) into embeddings**

Sense-Plan-Act AiO | Navigation with VLA



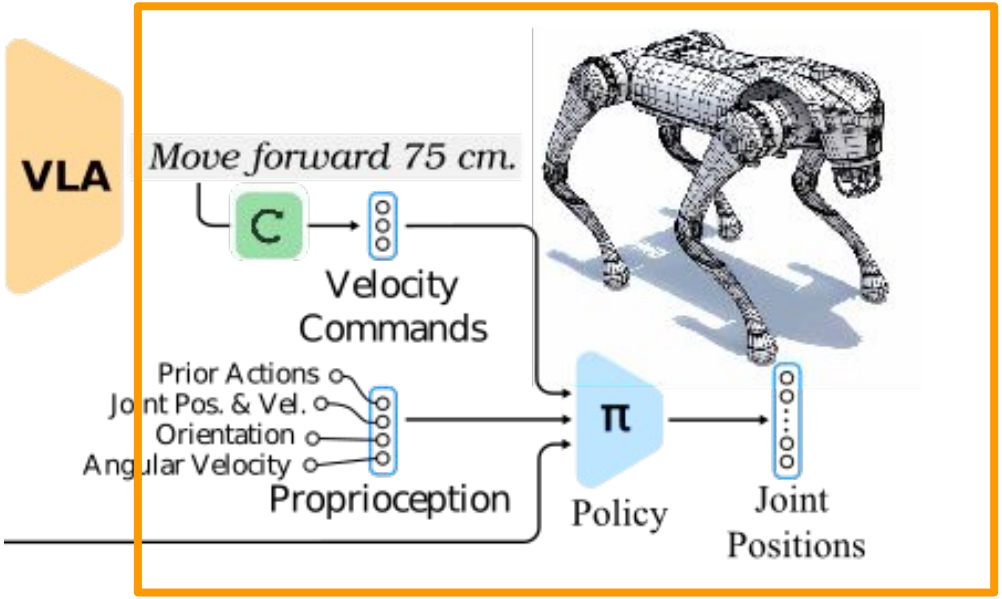
2. Build a **prompt for a LLM, including image embeddings** that are reprojected into text using a MLP

Sense-Plan-Act AiO | Navigation with VLA



3. Fine-tune (or instruct) LLM to **generate high-level instructions**

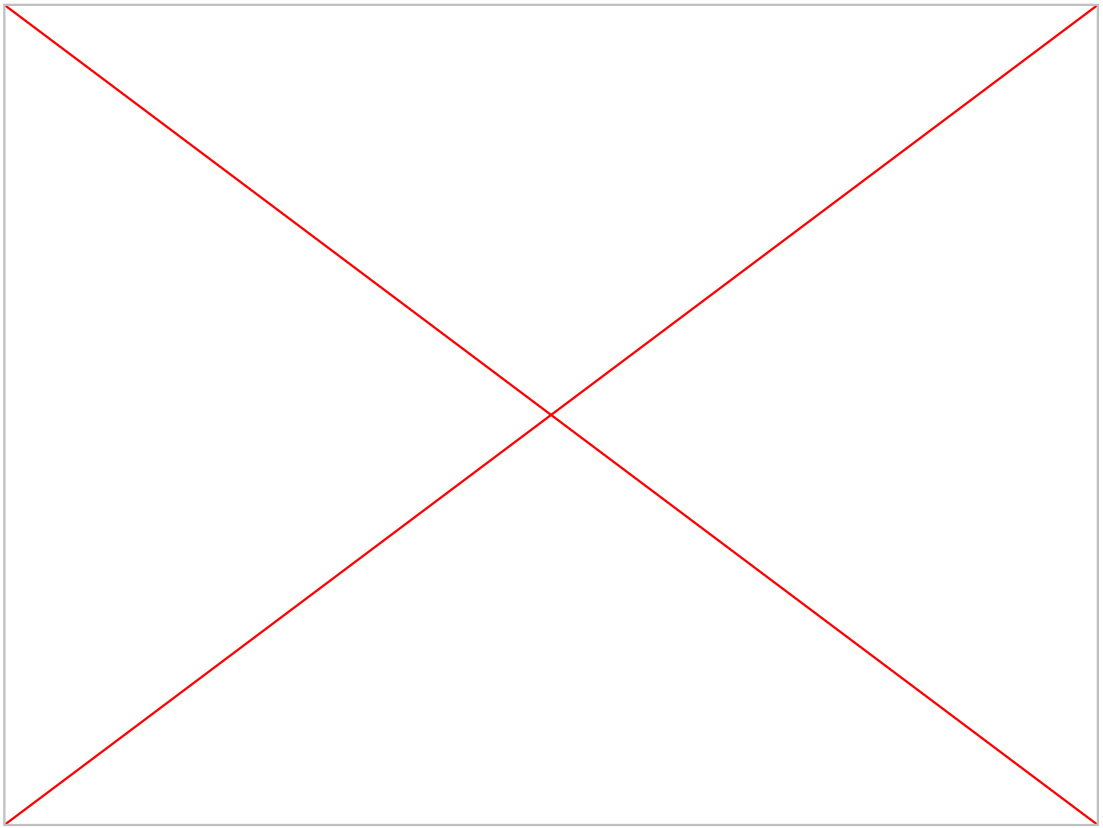
Sense-Plan-Act AiO | Navigation with VLA



4. **Map high-level actions into low-level commands** and feed a robot-specific control policy.

Sense-Plan-Act AiO | Navigation with VLA

This is the result:



Take Home Messages

- **High-level approach**: map input data into embeddings and let a transformer to learn dependencies and make predictions

- **Zero-Shot or Few Shot generalization**: after training/fine-tuning, it is possible to adapt to new scenario with no/few examples.

- **Multi-modal and Multi-task**: the same approach can be used for different robots, different tasks, different sensors

Limitations

- **Large Datasets**: pretraining of large models requires large datasets (millions of samples)
  - Dataset Creation and Curation
  - Synthetic Data and Simulations

- **Extreme Computations**: a lot of computational is needed to train and run large models and large memory is needed to store data
  - Do we really need very large models?

- **Out-of-Domain Generalization**: the zero/few-shot capabilities are limited to the domain described by training data
  - Problems in specific domain, such as industrial applications

References

Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017) - ***[slide 7-30]***

Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020). - ***[slide 13-15]***

Morrison, Douglas, Peter Corke, and Jürgen Leitner. "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach." arXiv preprint arXiv:1804.05172 (2018). - ***[slide 35-39]***

Fang, Hao-Shu, et al. "Graspnet-1billion: A large-scale benchmark for general object grasping." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. - ***[slide 35-39]***

Jin, Shiyu, et al. "Reasoning grasping via multimodal large language model." arXiv preprint arXiv:2402.06798 (2024). - ***[slide 40-45]***

M. Gou, H. -S. Fang, Z. Zhu, S. Xu, C. Wang and C. Lu, "RGB Matters: Learning 7-DoF Grasp Poses on Monocular RGBD Images," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021, pp. 13459-13466, doi: 10.1109/ICRA48506.2021.9561409. - ***[slide 35-39]***

Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. - ***[slide 33]***

Dalal, Murtaza, et al. "Neural mp: A generalist neural motion planner." arXiv preprint arXiv:2409.05864 (2024). - ***[slide 47-50]***

Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." arXiv preprint arXiv:2307.15818 (2023). - ***[slide 56-61]***

Brohan, Anthony, et al. "Rt-1: Robotics transformer for real-world control at scale." arXiv preprint arXiv:2212.06817 (2022). - ***[slide 56-61]***

Driess, Danny, et al. "Palm-e: An embodied multimodal language model." arXiv preprint arXiv:2303.03378 (2023). - ***[slide 56-61]***

Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2024). - ***[slide 40-45]***

Team, Octo Model, et al. "Octo: An open-source generalist robot policy." arXiv preprint arXiv:2405.12213 (2024). - ***[slide 62-66]***

Cheng, An-Chieh, et al. "Navila: Legged robot vision-language-action model for navigation." arXiv preprint arXiv:2412.04453 (2024). - ***[slide 67-72]***

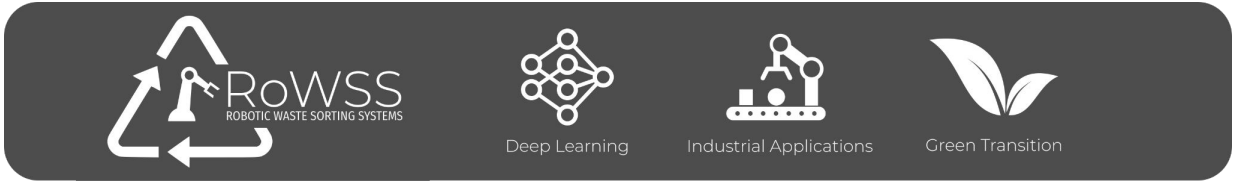
Barcellona, Leonardo, et al. "Fsg-net: a deep learning model for semantic robot grasping through few-shot learning." 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023. - ***[slide 35-39]***

Liang, Jacky, et al. "Code as policies: Language model programs for embodied control." 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023.- ***[slide 51-54]***

Chi, Cheng, et al. "Diffusion policy: Visuomotor policy learning via action diffusion." The International Journal of Robotics Research (2023): 02783649241273668.- ***[slide 55]***

Our Research and Thesis

Deep Learning Robotics and Green Transition



Introduction

Develop **AI and robotic technologies** to recover waste and reduce natural resources exploitation.

Introduction

Develop **AI and robotic technologies** to recover waste and reduce natural resources exploitation.

WHY

Boost a more Circular Economy

Make an Impact in the Real-World

Better Valorization of Human Labour

Introduction

Develop **AI and robotic technologies** to recover waste and reduce natural resources exploitation.

WHAT

Perception and Robotics

AI applied in Industry

Develop Cutting-Edge Technologies

Introduction

Develop **AI and robotic technologies** to recover waste and reduce natural resources exploitation.

HOW

IT+Robotics

大阪大学
OSAKA UNIVERSITY

SM competence center ACT

Research Projects

Industrial Projects

Business and International Collaboration

Research Areas

A1. Prototype Development

A2. Data Efficient Learning

A3. Deep Learning for HSI and MSI

A4. Semantic Grasping

A5. Advanced Manipulation

A1. Prototype Development

Developing a real-world robotic waste sorting system

Develop software for real hardware and test it with experiments.

A1 | Mass Estimation from RGB-D Images

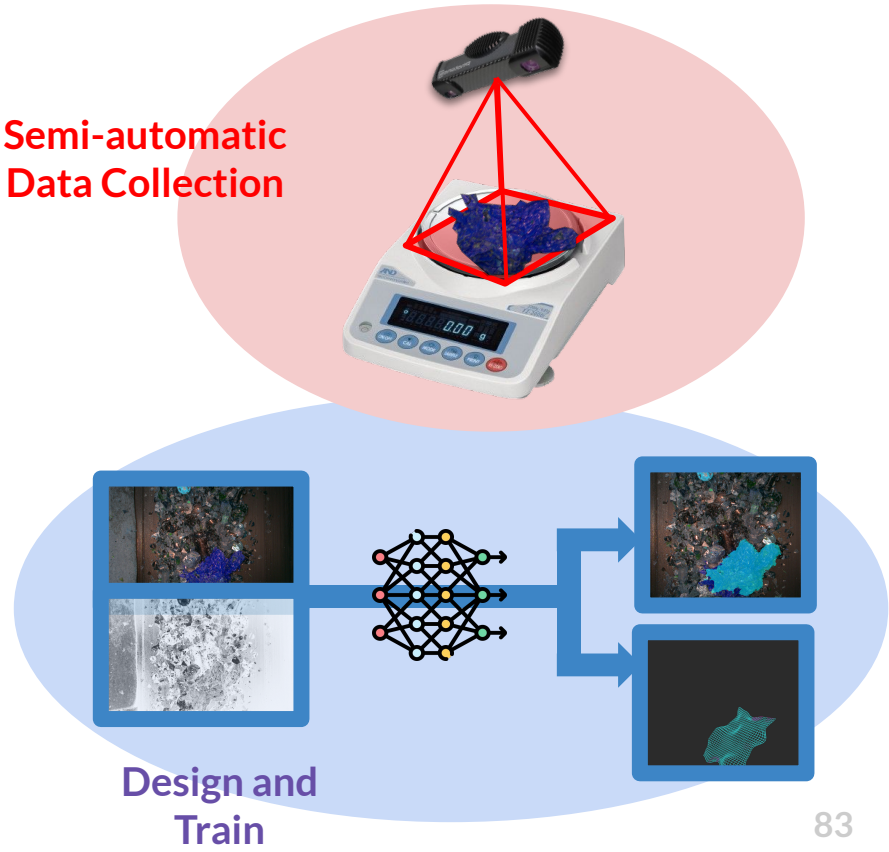
PROJECT OUTLINES

Description
Develop a Deep Learning method for mass estimation of contaminants in waste streams.

Impact
- Tackle a **critical open challenge** such mass estimation
- Learn to **collect data**
- Learn to **design novel DL model**

Operational Plan
- Literature search on volume/mass estimation from RGB-D
- Setup of the vision and data collection system (camera + scale)
- Semi-automatic data collection
- Build and train the model
- Test the model

Requirements
- Basic knowledge of DL and Pytorch
- Familiarity with computer vision tasks like object detection and semantic segmentation



A1 | Camera Synchronization

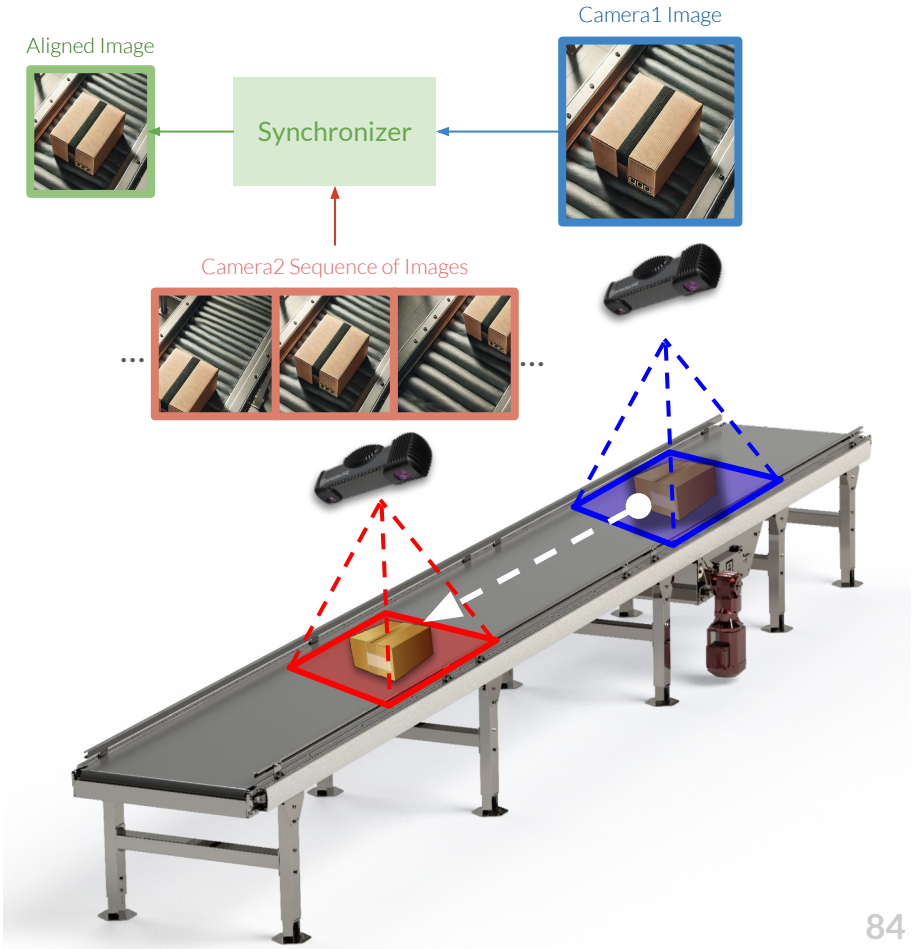
PROJECT OUTLINES

Description
Camera synchronization is crucial for multi-camera setup, but in absence of encoders is very complex

Impact
- Tackle the **real-world challenge** of camera synchronization without encoder
- Developing learning-based algorithms
- Learn to use **industrial systems**

Operational Plan
- Literature search on camera synchronization
- Setup of the dual-camera
- Delevelop method for image synchronization (e.g. data acquisition with tags and feature matching like MatchAnything)
- Test the method

Requirements
- Basic knowledge Python and C++
- Familiarity with computer vision

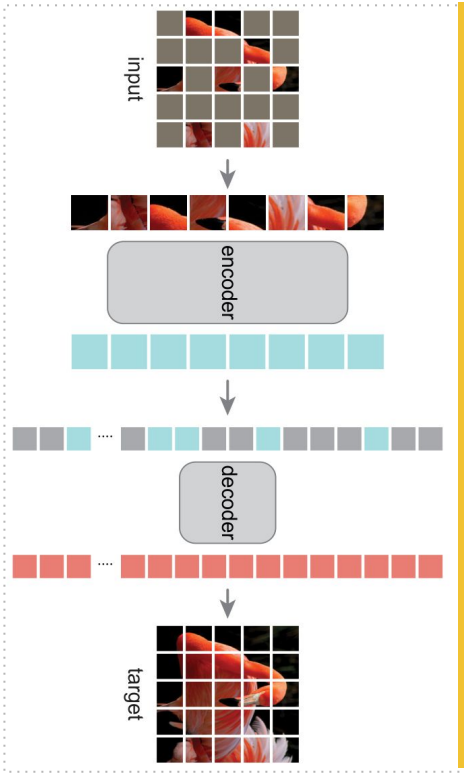


A2. Data Efficient Learning

Reduce dependency on manual annotations for training and fine-tuning DL models

Acquire and label data can be expensive and time consuming, reducing applicability and flexibility

SYNTHETIC DATA & AUGMENTATION



WEAK- & SELF-SUPERVISED LEARNING

A2 | Augmented Synthetic Data Generation

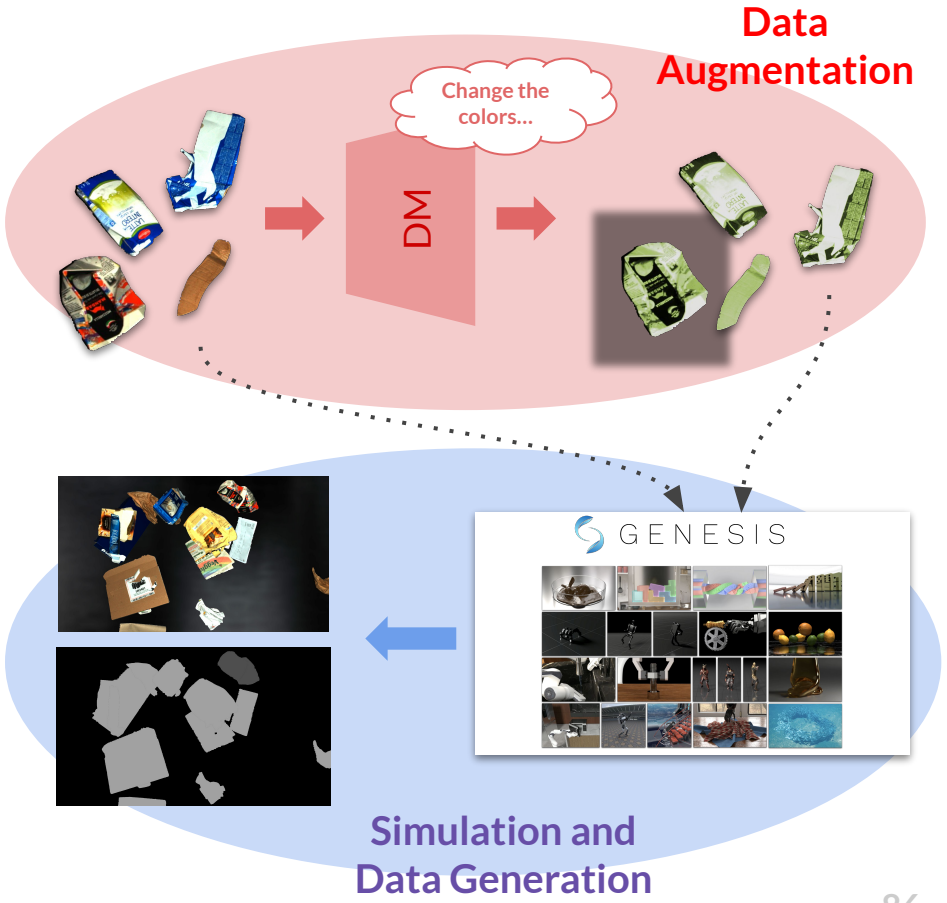
PROJECT OUTLINES

Description
Synthetic data generation using Diffusion Models (DM) and physics simulators.

Impact
- Learn to cutting edge methods, such as **DM**
- Learn to **use physics simulator**
- Extend applicability of **DL in real-world**

Operational Plan
- Literature search on DM for augmentation
- Integration of a DM to apply augmentations to (dataset) objects
- Setup the simulation environment
- Synthetic and augmented data generation

Requirements
- Basic knowledge of DL and Pytorch
- Knowledge of Python and Computer Vision



A2 | Latent Data Augmentation with DM

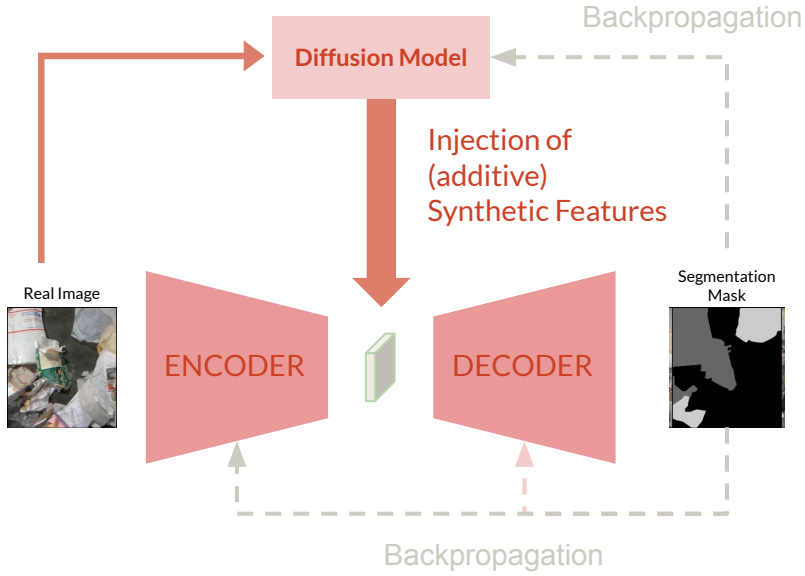
PROJECT OUTLINES

Description
Generate Augmented Latent features in place of synthetic Images to train models

Impact
- Learn to **use DM**
- Learn to **train DL models**
- Extend applicability of **DL in real-world**

Operational Plan
- Literature search on DM for augmentation and latent feature augmentation
- Integration of a DM to train a DL model for classification and/or semantic segmentation
- Design of a joint training loop (DA only in training) or joint train-test loop (DA also at test time)
- Train and test on public datasets

Requirements
- Basic knowledge of DL and Pytorch
- Familiarity with computer vision tasks like object detection and semantic segmentation



A2 | Few-Shot Continual Learning

PROJECT OUTLINES

Description

The goal is to enable deep learning models to extend and update their knowledge dynamically through data-efficient, human-like learning from a few examples, rather than relying solely on static knowledge fixed during training.

Impact

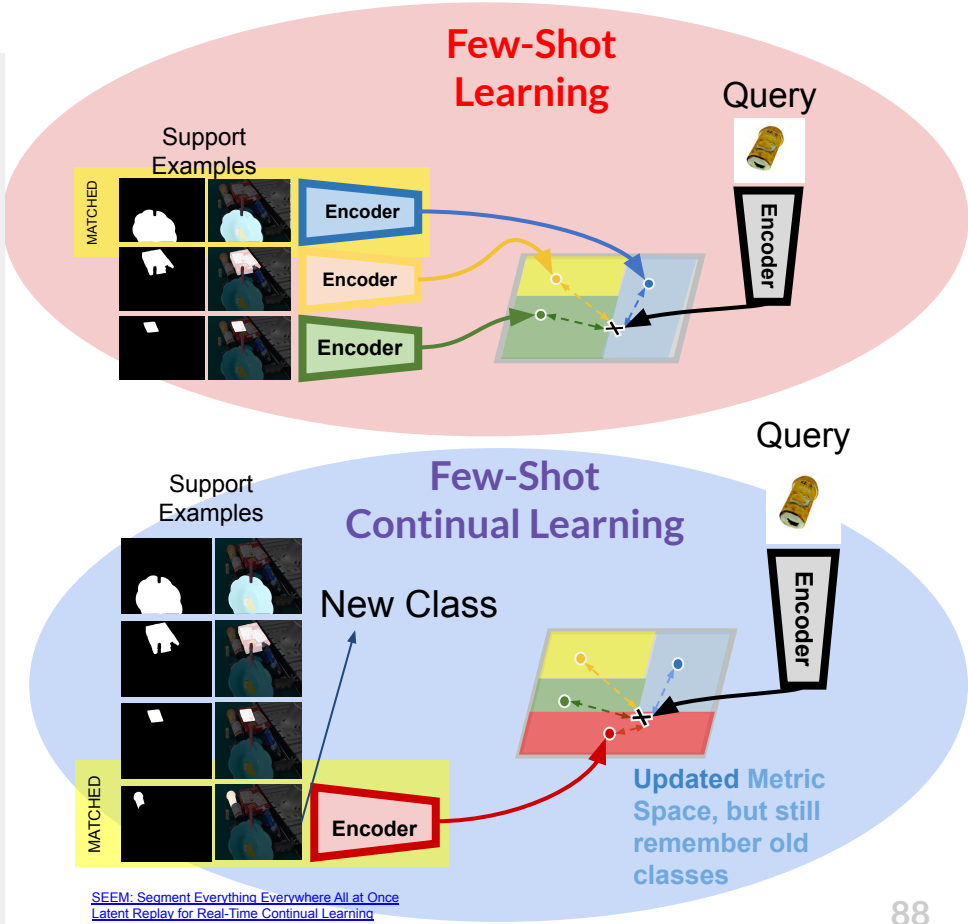
- Learn to **Continual Learning methods**
- Learn to **Few-Shot methods**

Operational Plan

- Literature Review on Continual and Few-Shot Learning
- Design a suitable representation to store the knowledge (e.g. a model that maps images in a metric space)
- Design strategy to populate the metric space and use it at inference time

Requirements

- Good knowledge of DL and Pytorch
- Familiarity with computer vision tasks like object detection and semantic segmentation



A2 | RLHF for Vision and Robotics

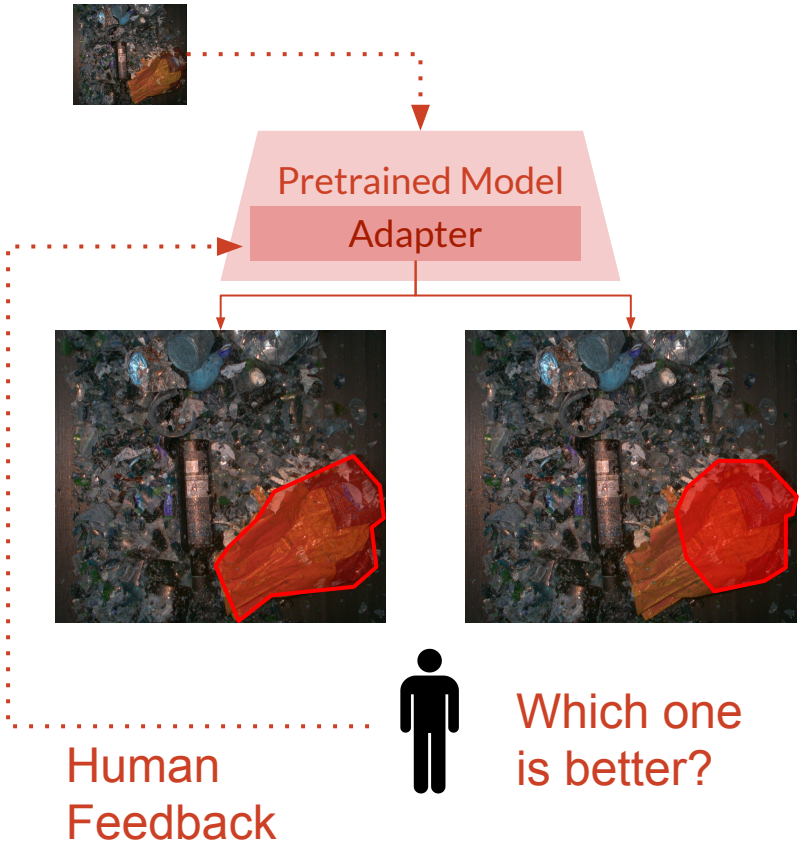
PROJECT OUTLINES

Description
Develop Reinforcement Learning from Human Feedback (RLHF) approaches to fine-tune pre-trained models for vision in novel domains.

Impact
- Learn cutting-edge methods, such as **RLHF**
- Extend applicability of **DL in real-world**

Operational Plan
- Literature search on RLHF
- Proposal of a RLHF approach for vision or robotics (like grasping)
- Implementation and training of the approach
- Train and test on datasets and real-world data

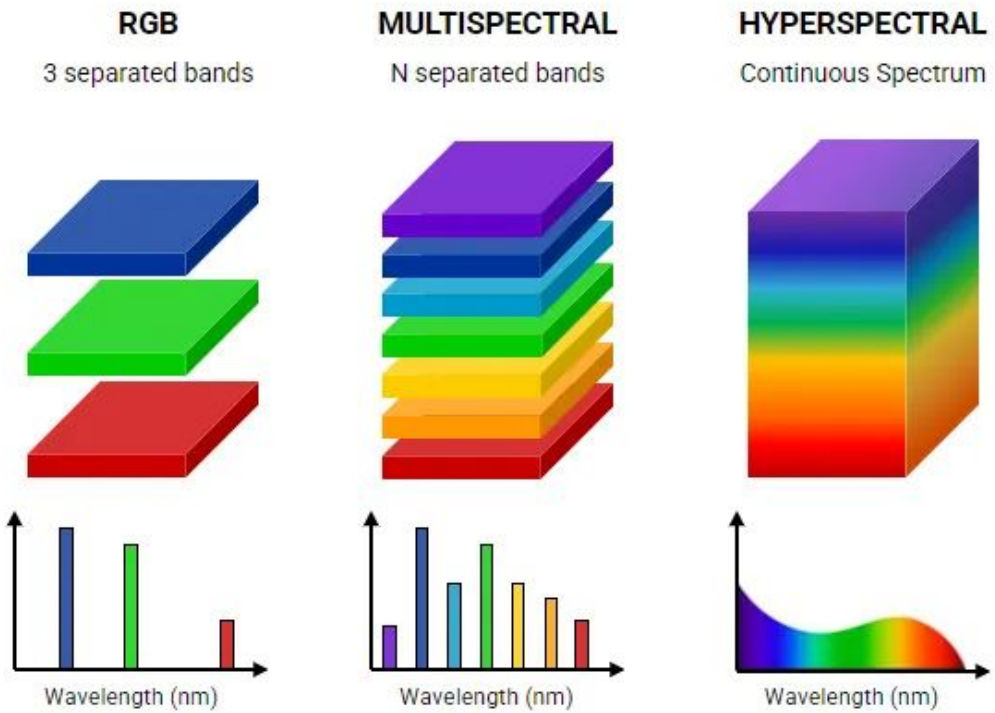
Requirements
- Good knowledge of DL and Pytorch
- Familiarity with computer vision and robotics



A3. Deep Learning for HSI & MSI

HSI and MSI technologies go beyond human vision

But we do not have large datasets and DL architecture to efficiently exploit these data



A3 | Model Ensembling for HSI & MSI

PROJECT OUTLINES

Description

The idea is to leverage the knowledge of pre-trained RGB models to efficiently process HSI and MSI data, since features are similar.

Impact

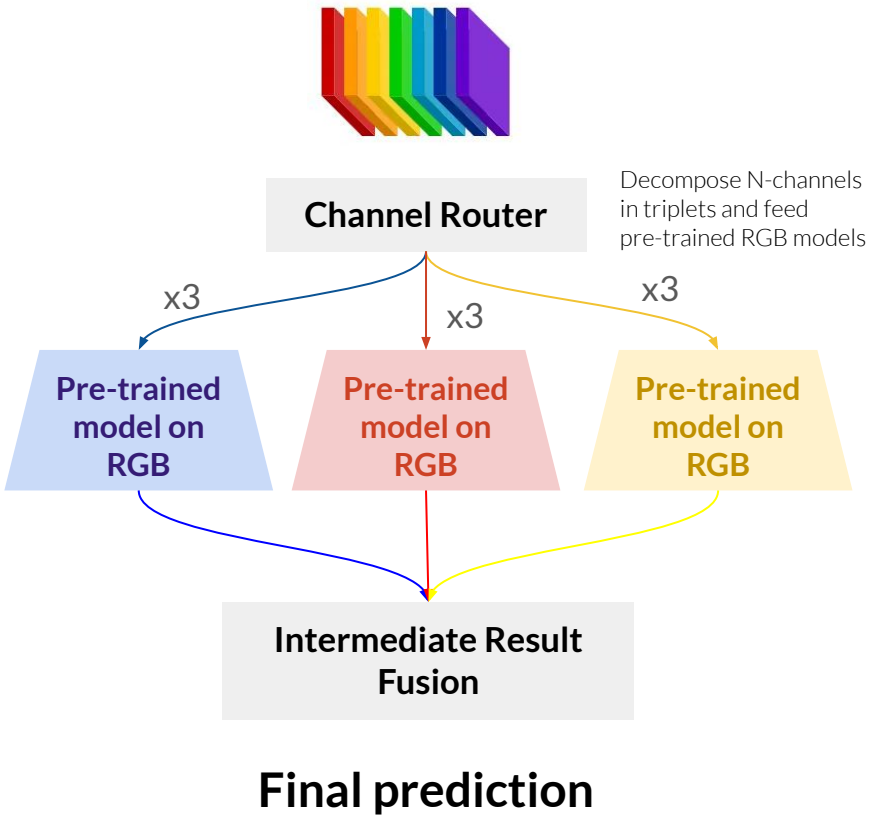
- **Pushing the limits of computer vision**, akin to the leap from black-and-white to color imaging.
- Open applicability of **HSI & MSI for complex tasks**

Operational Plan

- Literature search on HSI&MSI and Model Ensembling
- Subsampling and routing of HSI into 3-channel sub image. Start with a uniform sampling, we can learn a dynamic router in future (like in MoE)
- Implementation of model ensembling by fusing the output of intermediate models

Requirements

- Good knowledge of DL and Pytorch
- Familiarity with computer vision



A3 | Few-Shot Learning for HSI & MSI

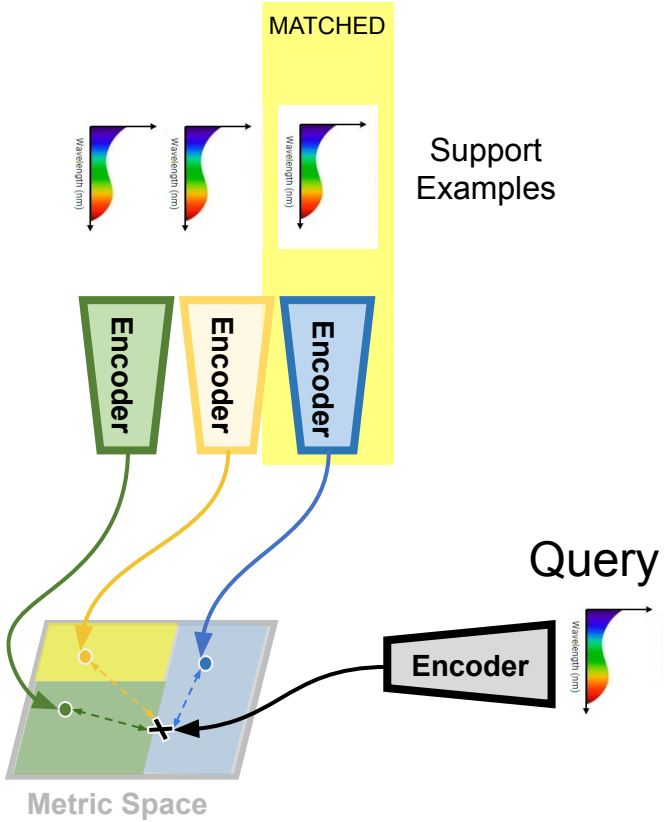
PROJECT OUTLINES

Description
Adapt few-shot methods to HSI/MSI, exploiting the channel redundancy.

Impact
- **Pushing the limits of computer vision**, akin to the leap from black-and-white to color imaging.
- Open applicability of **HSI & MSI for complex tasks**

Operational Plan
- Literature search on HSI&MSI and Few-Shot Learning
- Encoder Training: train a prototype encoder, able to exploit channel redundancy (i.e. each pixel is a vector of N channels)

Requirements
- Good knowledge of DL and Pytorch
- Familiarity with computer vision



A4. Semantic Grasping

Recognize and predict a grasping pose for a target object or class

keeping a data-efficient approach and real-time constraints

L. Barcellona, A. Bacchin, A. Gottardi, E. Menegatti and S. Ghidoni, "FSG-Net: a Deep Learning model for Semantic Robot Grasping through Few-Shot Learning," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, United Kingdom, 2023, pp. 1793-1799, doi: 10.1109/ICRA48891.2023.10160618 .

A4 | Multi-Modal Few-Shot Semantic Grasping

PROJECT OUTLINES

Description

Few-Shot Grasping involves identifying objects or categories from limited examples and computing grasp poses. The approach integrates visual-language models to enhance the process.

Impact

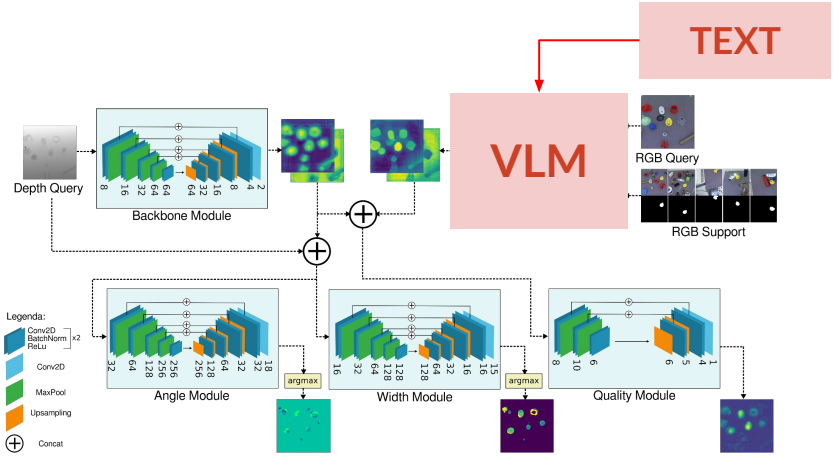
- Learning to use **VLM for robotics**
- Increase real-world applicability of robots

Operational Plan

- Literature review on VLM for grasping and Few-Shot learning
- Extend the Few-Shot Module with a VLM (keeping in mind rea-time)
- Train and test the grasping pipeline

Requirements

- Good knowledge of DL and Pytorch
- Familiarity with computer vision and robotics



Support Set is Text + Images

A4 | 6DoF Semantic Grasping

PROJECT OUTLINES

Description

Few-Shot Grasping involves identifying objects or categories from limited examples and computing grasp poses. We want to extend it for 6DOF poses.

Impact

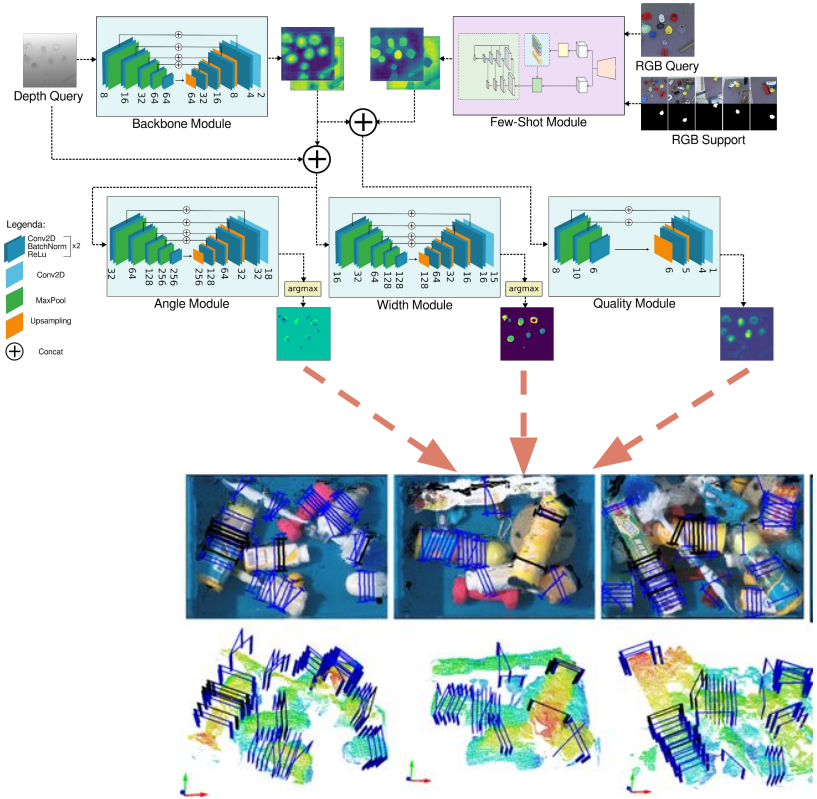
- Learning to **design DL networks**
- Increase real-world applicability of robots

Operational Plan

- Literature review on 6DOF grasping
- Definition of a strategy from a single-view to multi-view/3D representation
- Synthesis of a 6DOF grasping pose with semantics

Requirements

- Good knowledge of DL and Pytorch
- Familiarity with computer vision and robotics



A4 | Semantic Grasping using Synthetic Depth

PROJECT OUTLINES

Description

3D data is challenging in cluttered, dynamic scenarios. We aim to optimize grasping poses for target objects using only RGB images and monocular depth estimators, starting with vacuum grippers.

Impact

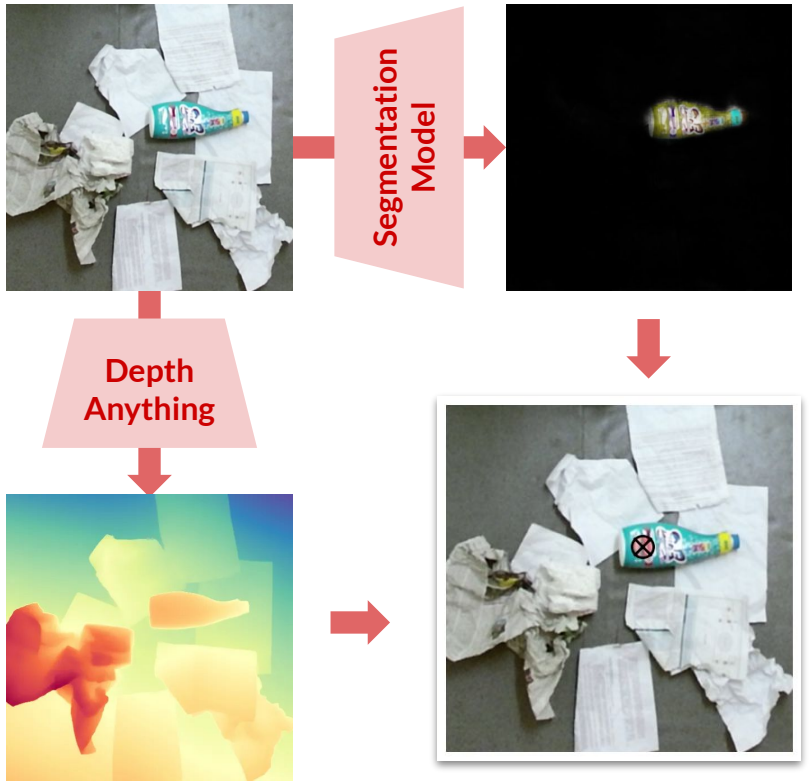
- Learning to **design DL networks**
- Learn **modern Depth Estimator** like Depth Anything

Operational Plan

- Literature review on robotic grasping and monocular depth estimators
- Definition of a strategy, e.g. combining semantic segmentation with Depth Anything,
- Train and test the model (also on real robot if possible)

Requirements

- Basic knowledge of DL and Pytorch
- Familiarity with computer vision, 3D data and robotics



A4 | Self-supervised Semantic Grasping

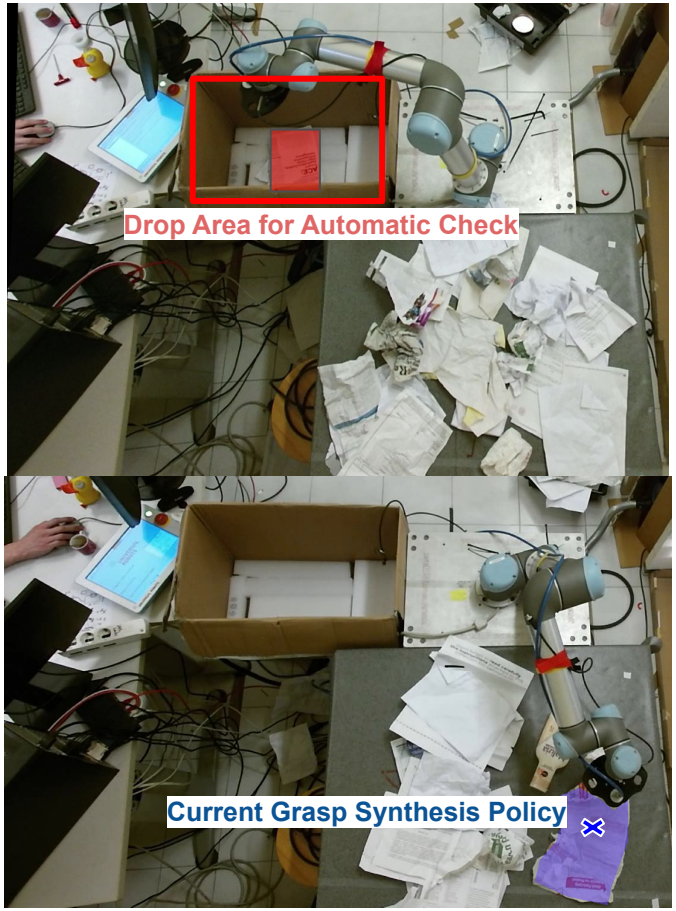
PROJECT OUTLINES

Description
Train a model to compute a grasping point, given a segmentation mask, using self-supervised learning.

Impact
- Learning to **design DL networks**
- Learn **to deal with self-supervised learning** applied to robotics

Operational Plan
- Literature review on robotic grasping and self-supervised learning
- Developing of the self-supervision using visual feedback → start from a pre-trained model or train in simulation first
- Implementation and training of the model for semantic-aware grasp synthesis

Requirements
- Basic knowledge of DL and Pytorch
- Familiarity with computer vision and robotics



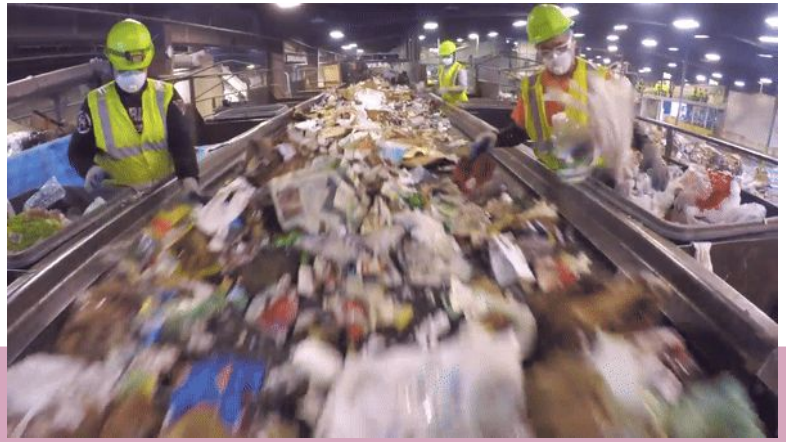
Did the robot pick the item?
Yes → Positive Feedback
No → Negative Feedback

Update Grasp Synthesis Model

A5. Advanced Manipulation

Enabling robots to manipulate cluttered, unstructured and variated objects

Needs of pre-grasp motion, like rummaging or preparing, and long-term reasoning for human-like manipulation



FROM THIS…

…TO THIS

A5 | Semantic Grasping in Clutters

PROJECT OUTLINES

Description

In cluttered scenes, target object may be not visible or partially visible. Introduction of scene reasoning and propaedeutic motion for grasping.

Impact

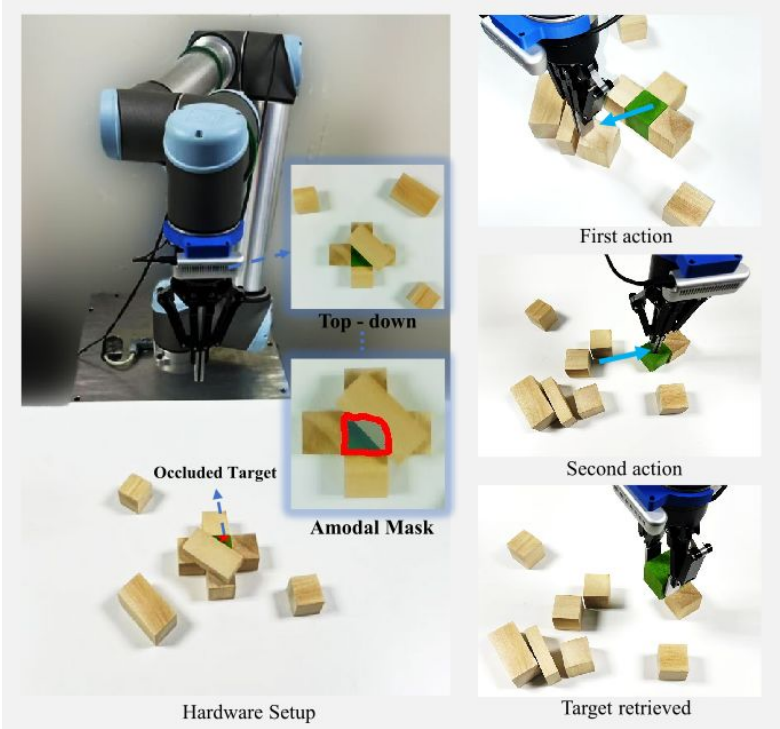
- Learning to **design complex DL networks**
- Usage of **VLM for scene reasoning**
- Enable application of robotics in unstructured environments

Operational Plan

- Literature review on grasping in cluttered scenes and VLM
- Developing of a modular DL pipeline for grasping in cluttered scenes → usage of simulations and real-word data
- Training and test

Requirements

- Good knowledge of DL and Pytorch
- Familiarity with computer vision, 3D data and robotics



H. Ding, Y. Zeng, Z. Wan and H. Cheng, "OPG-Policy: Occluded Push-Grasp Policy Learning with Amodal Segmentation," *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, United Arab Emirates, 2024, pp. 7257-7263, doi: 10.1109/IROS58592.2024.10802573

A5 | Semantic Grasp and Throw

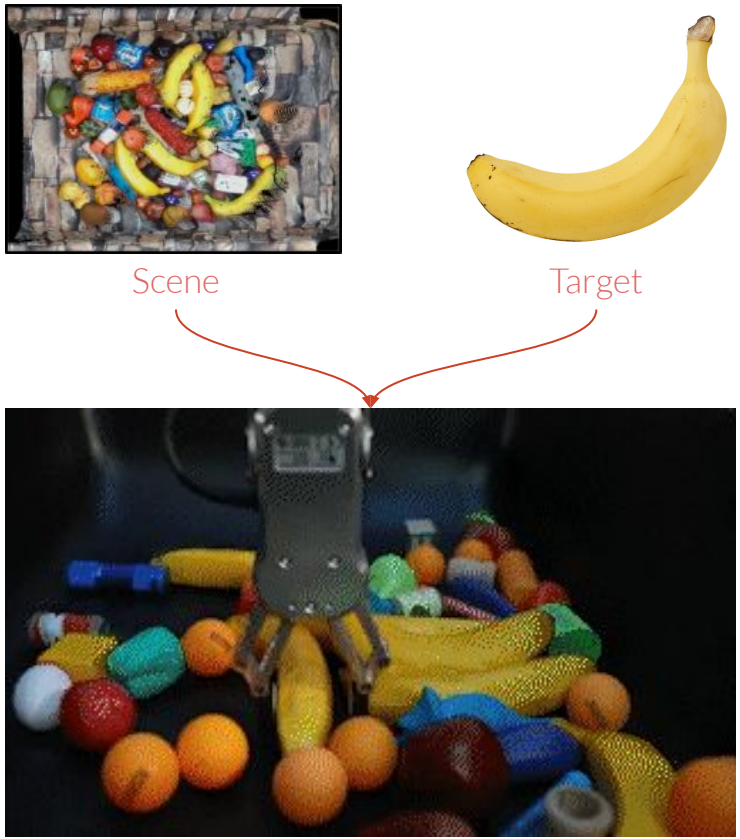
PROJECT OUTLINES

Description
When the fast manipulation is need, pick-and-throw can be more efficient. We want to extend semantic grasping with throwing actions.

Impact
- Learning to **design complex DL networks**
- Usage of **simulation and self-supervision**
- Enable application of robotics in unstructured environments

Operational Plan
- Literature review on grasping and throwing
- Developing of a unified DL pipeline to grasp and throw a target object
- Training and test

Requirements
- Good knowledge of DL and Pytorch
- Familiarity with computer vision, 3D data and robotics



Call to Action!

If you are interested in these (or related) topic,
get in touch with us!

Prof. Emanuele Menegatti

emanuele.menegatti@unipd.it

Alberto Bacchin

bacchinalb@dei.unipd.it