Physical Intelligence π

Short and Long-Term Memory as an Essential Ingredient for Long-Horizon Manipulation

MEM: Multi-Scale Embodied Memory for Vision Language Action Models

Marcel Torne, RSS 2026


autonomous, 4x, 00:00
π
autonomous, 2x, 00:00
π
autonomous, 4x, 00:00
π
CRAFTSMAN
autonomous, 2x, 00:00
π


FH700


405
CABINETS

What are we missing to solve these long-horizon tasks?

1. Keep track of task progress
2. Keep track of time
3. Reliable of the dexterity
4. Capacity to adapt in-context
5. Remember what has been tried and done before
6. …

Claim: We need memory to solve long-horizon tasks

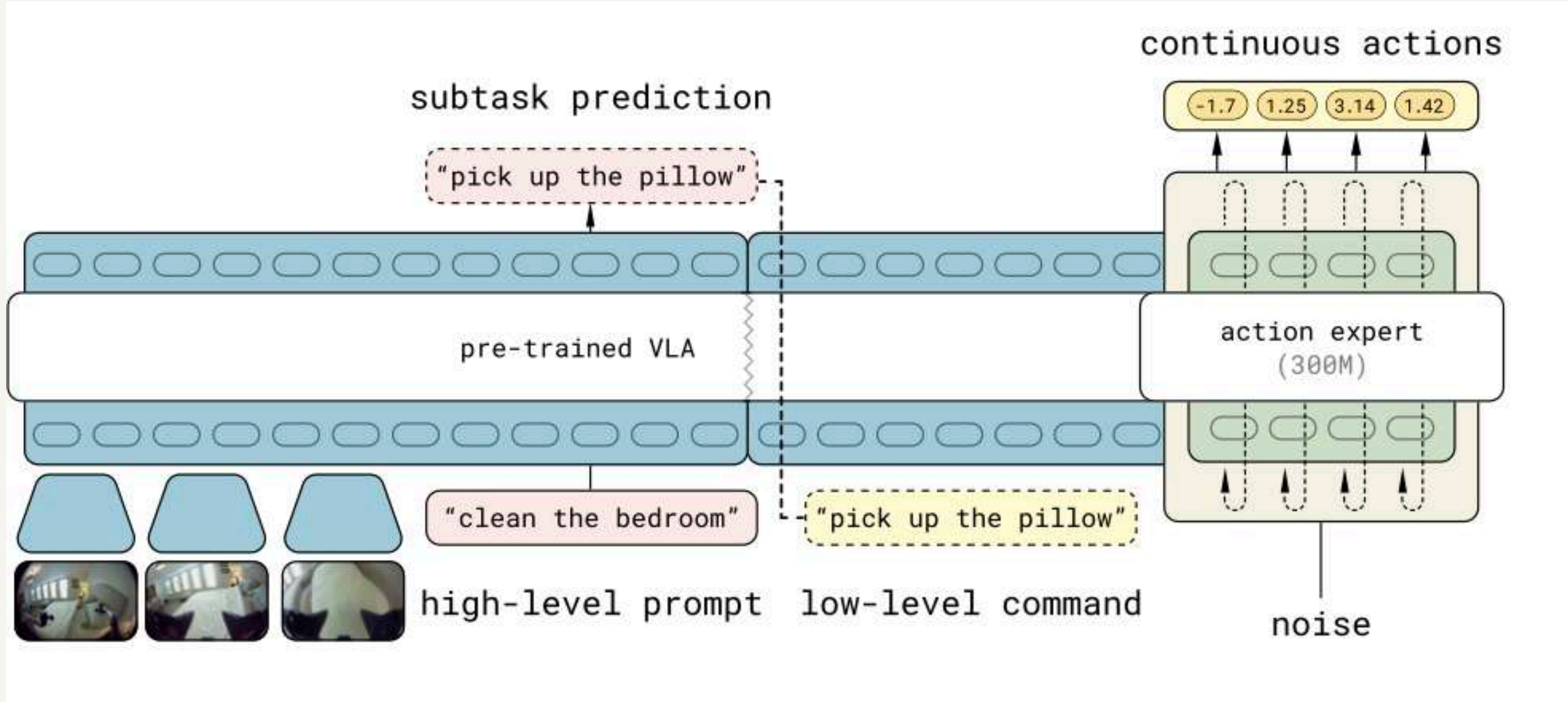
A lot of robot tasks require memory

Preparing the ingredients to cook a recipe

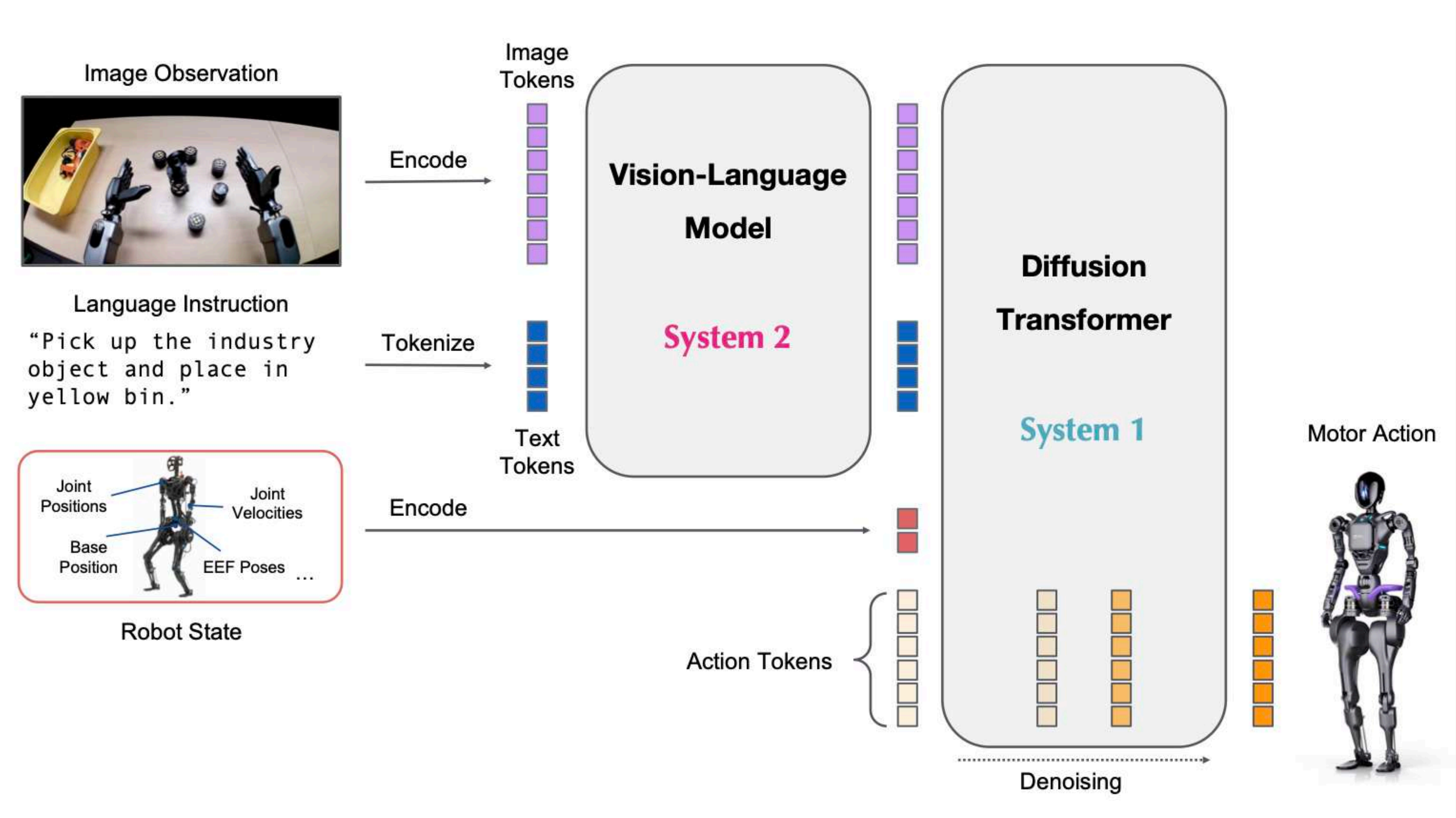
Unpacking the items in a grocery bag

However most robot policies don't have memory

Pi 0.5



GR00T



Why don’t robot policies use memory?   It is hard!

Two main problems:

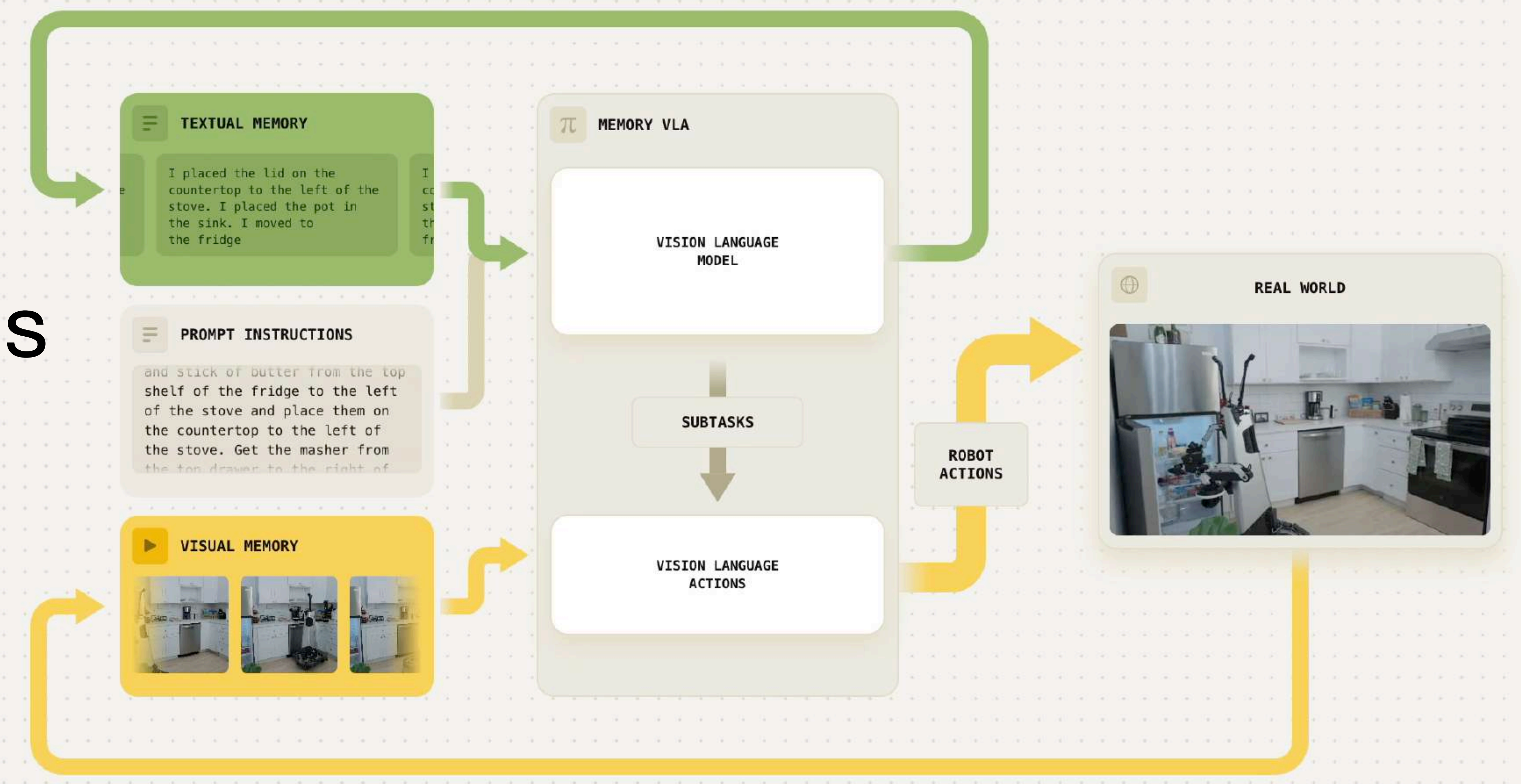
- Effectiveness
  - Policies perform worse because of distribution shifts and lack of data
- Efficiency
  - With longer context policies become more resource intensive
    - Longer to train and to do inference

*Learning Long-Context Diffusion Policies with Past Token Prediction, **Torne**** *et al, 2025*

Our solution? Compression!

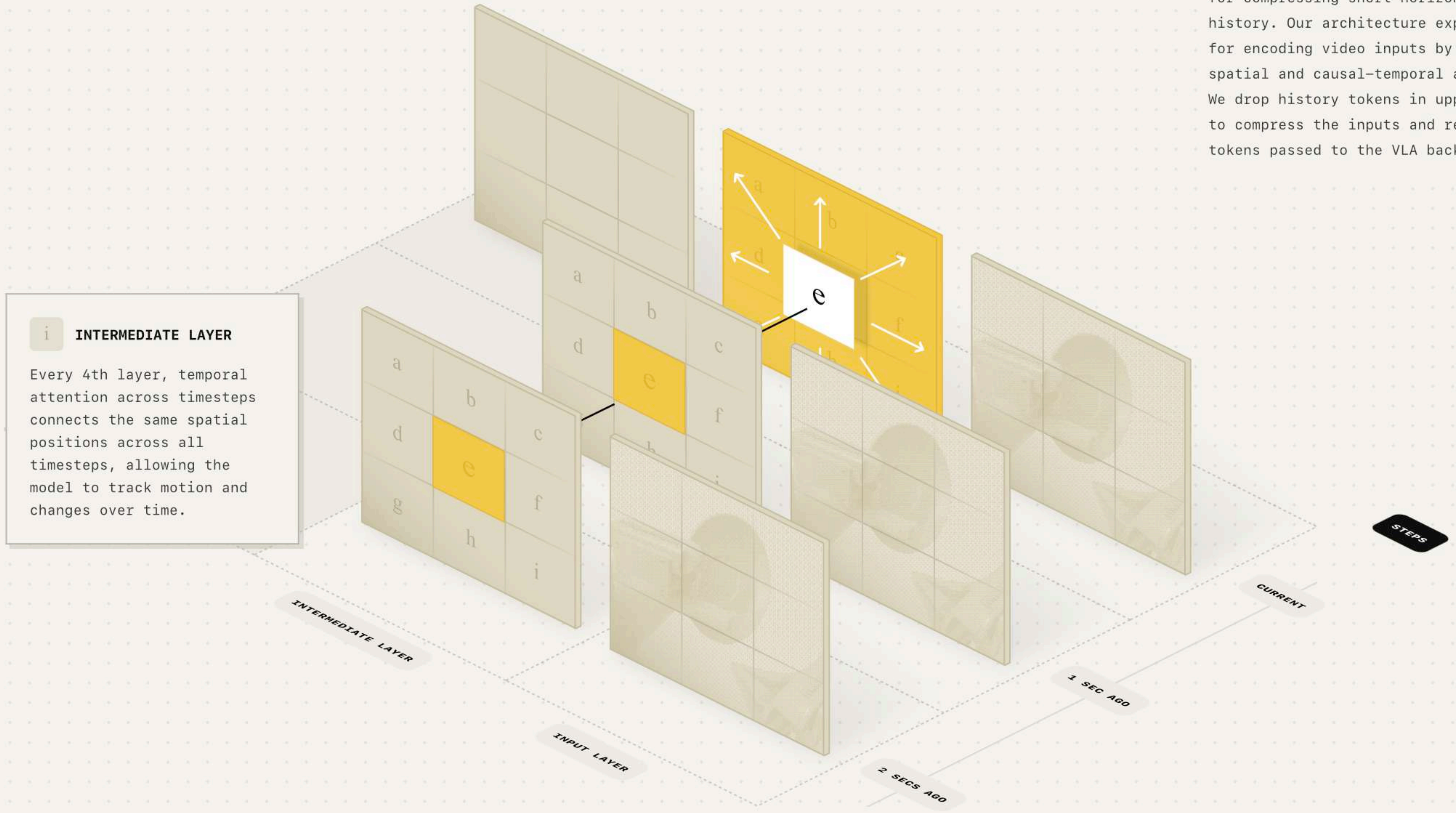
Decompose memory into two different types

- Short-context
  - Dense frames from the last few seconds needed for dexterous manipulation
- Long-context
  - Compressed language representation of the last minutes for general reasoning

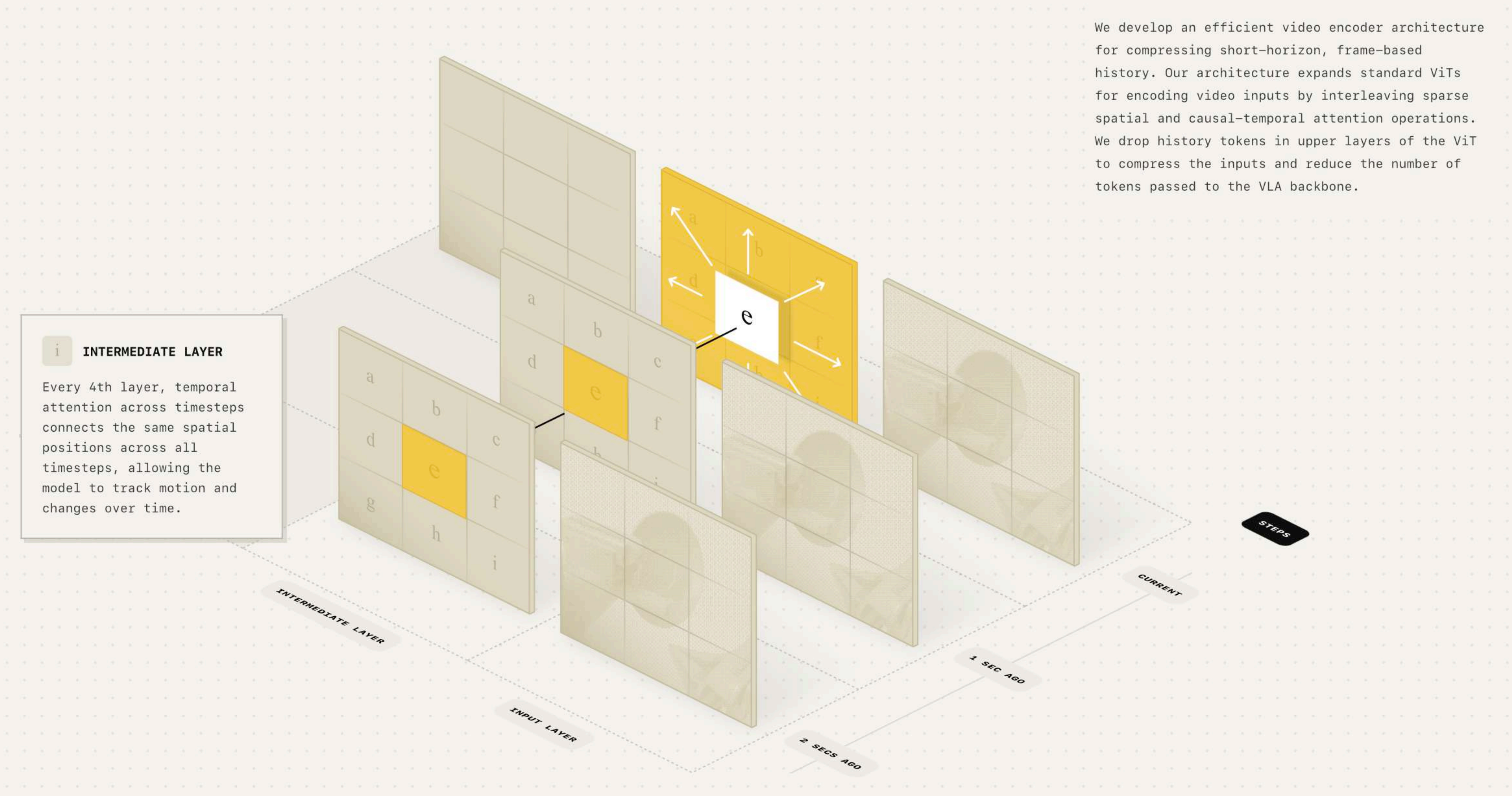


Short-Term Visual Memory: Method

We develop an efficient video encoder architecture for compressing short-horizon, frame-based history. Our architecture expands standard ViTs for encoding video inputs by interleaving sparse spatial and causal-temporal attention operations. We drop history tokens in upper layers of the ViT to compress the inputs and reduce the number of tokens passed to the VLA backbone.



Short-Term Visual Memory: Why this encoder?

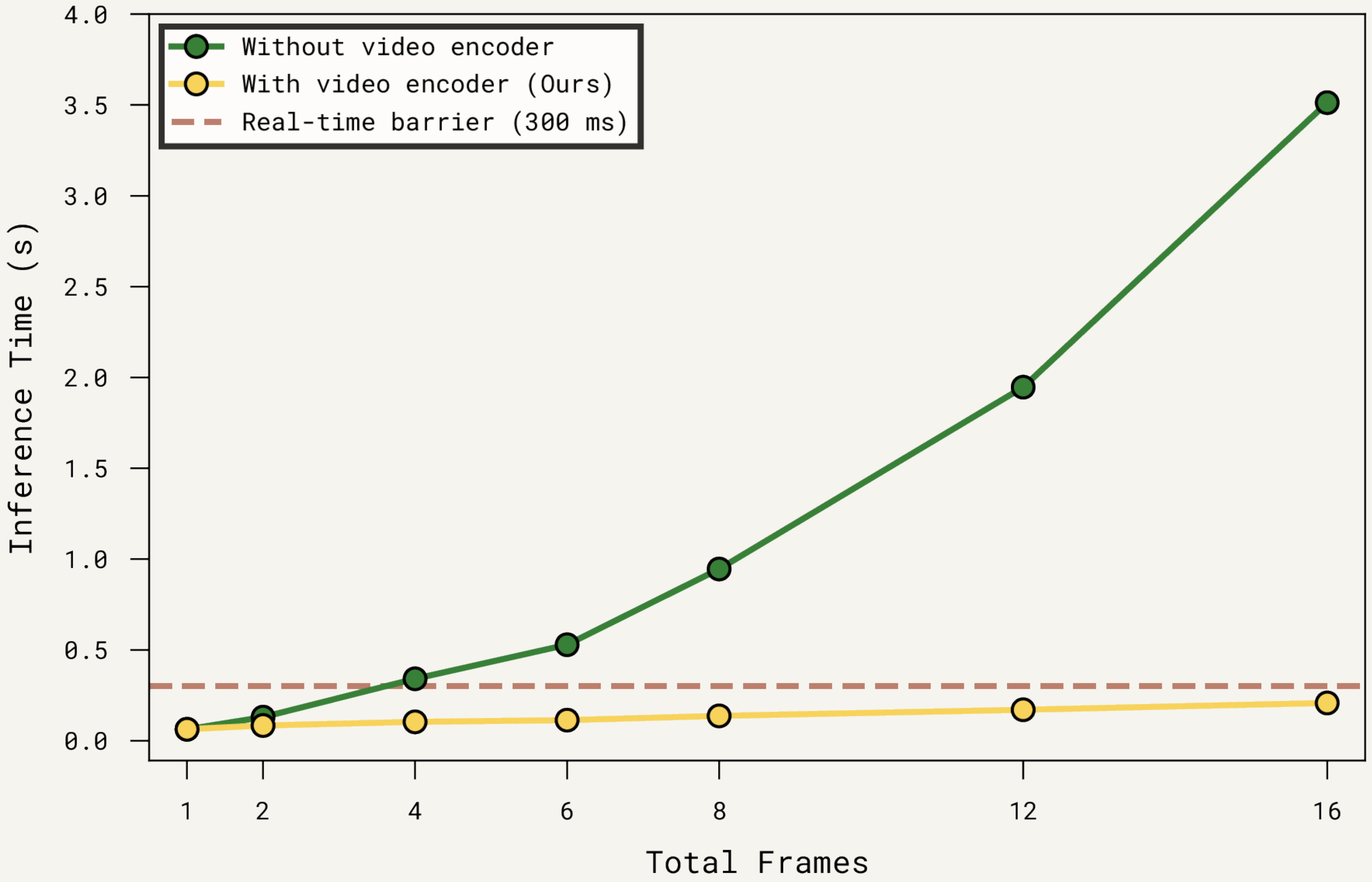


Why this architecture?

- Easy ViT initialization (only temporal attention is new)
- Compression of image sequence
- Fast inference

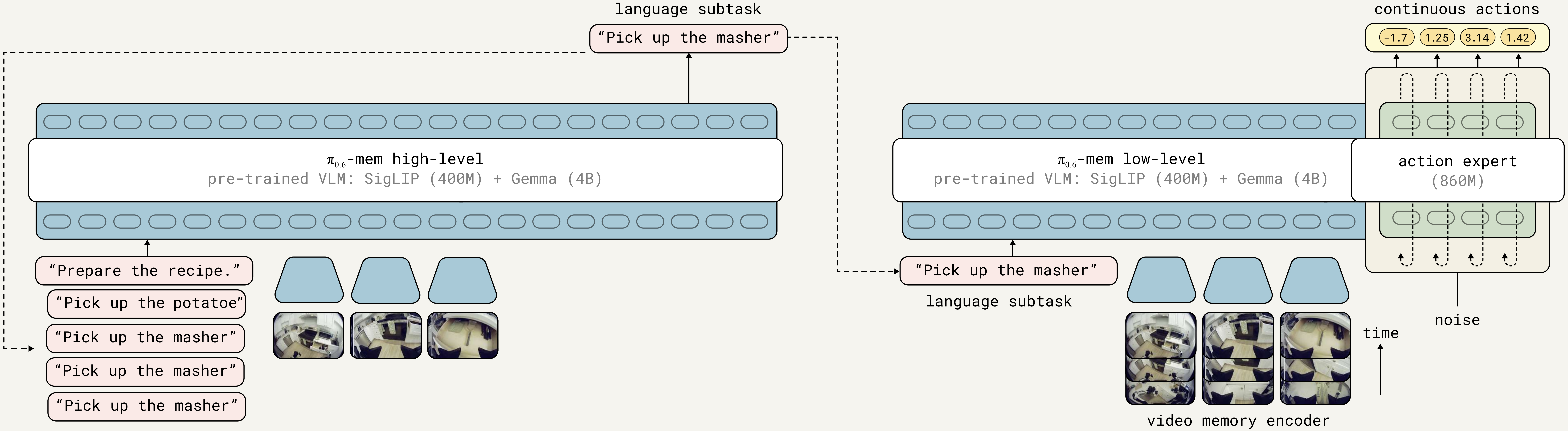
Short-term memory can solve a new category of tasks

Short-Term Visual Memory: Important Details



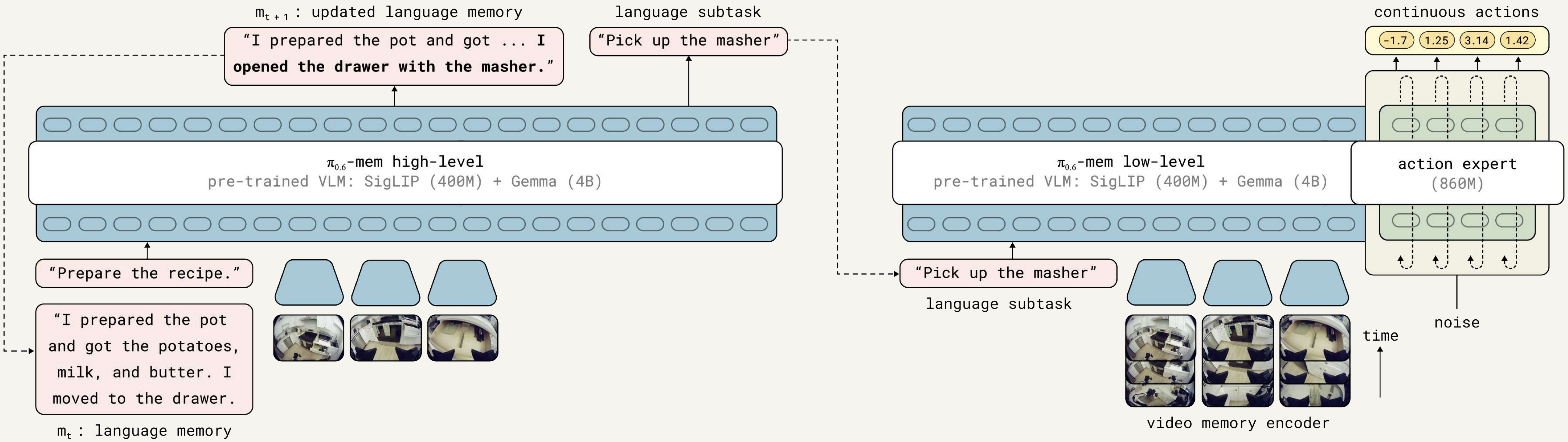
Our MEM encoder reduces the inference latency

Long-Term Textual Memory: Naive Textual



- Much more compressed representation than images (text uses less tokens)
- Not physically accurate
- Prone to distribution shifts

Long-Term Textual Memory: MEM (Ours)

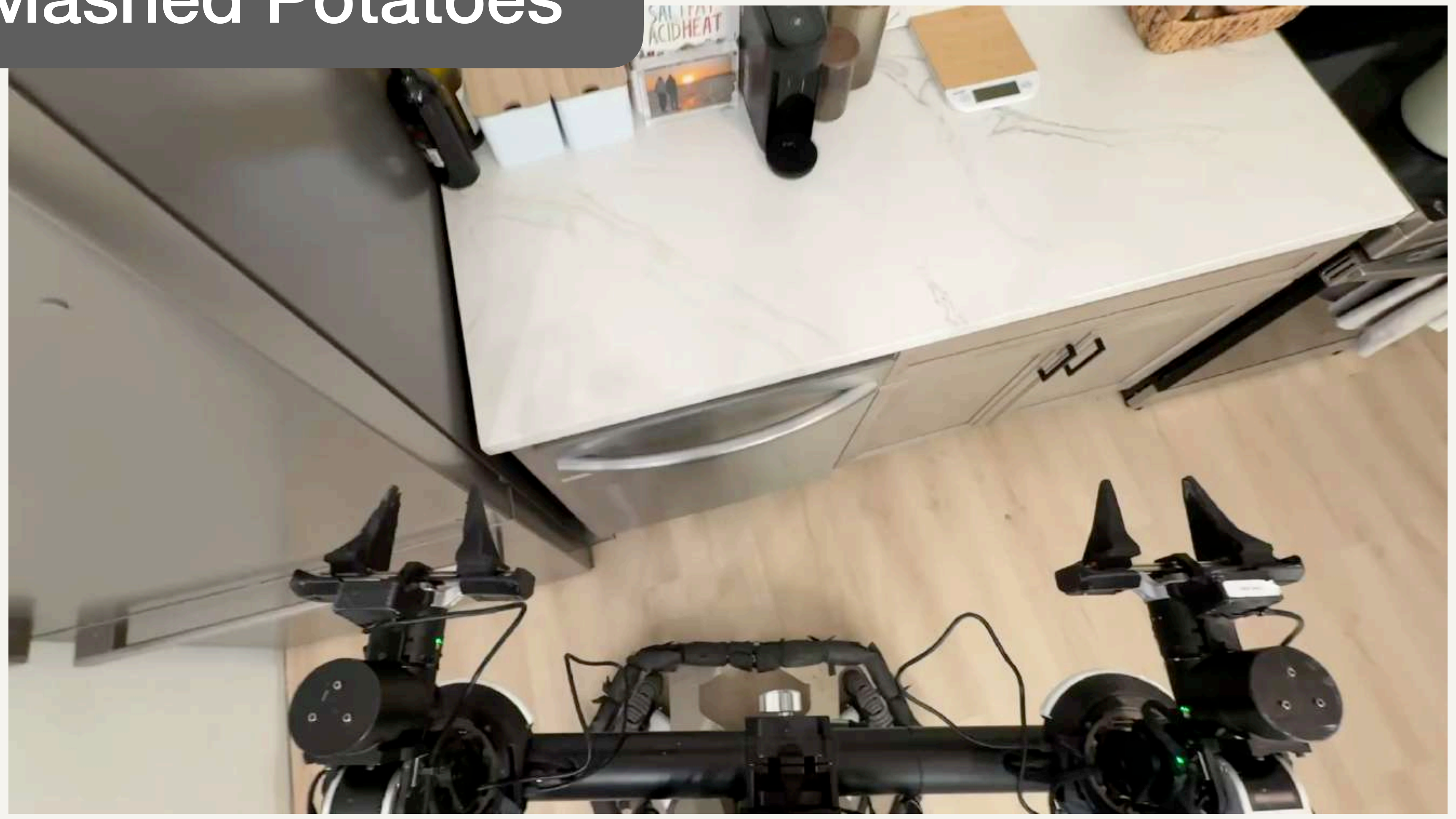


- Much more compressed representation than images (text uses less tokens)
- More physically accurate (the policy will update the memory with the physical events that did occurred)
- Less prone to distribution shifts (no update of the memory when nothing happens)

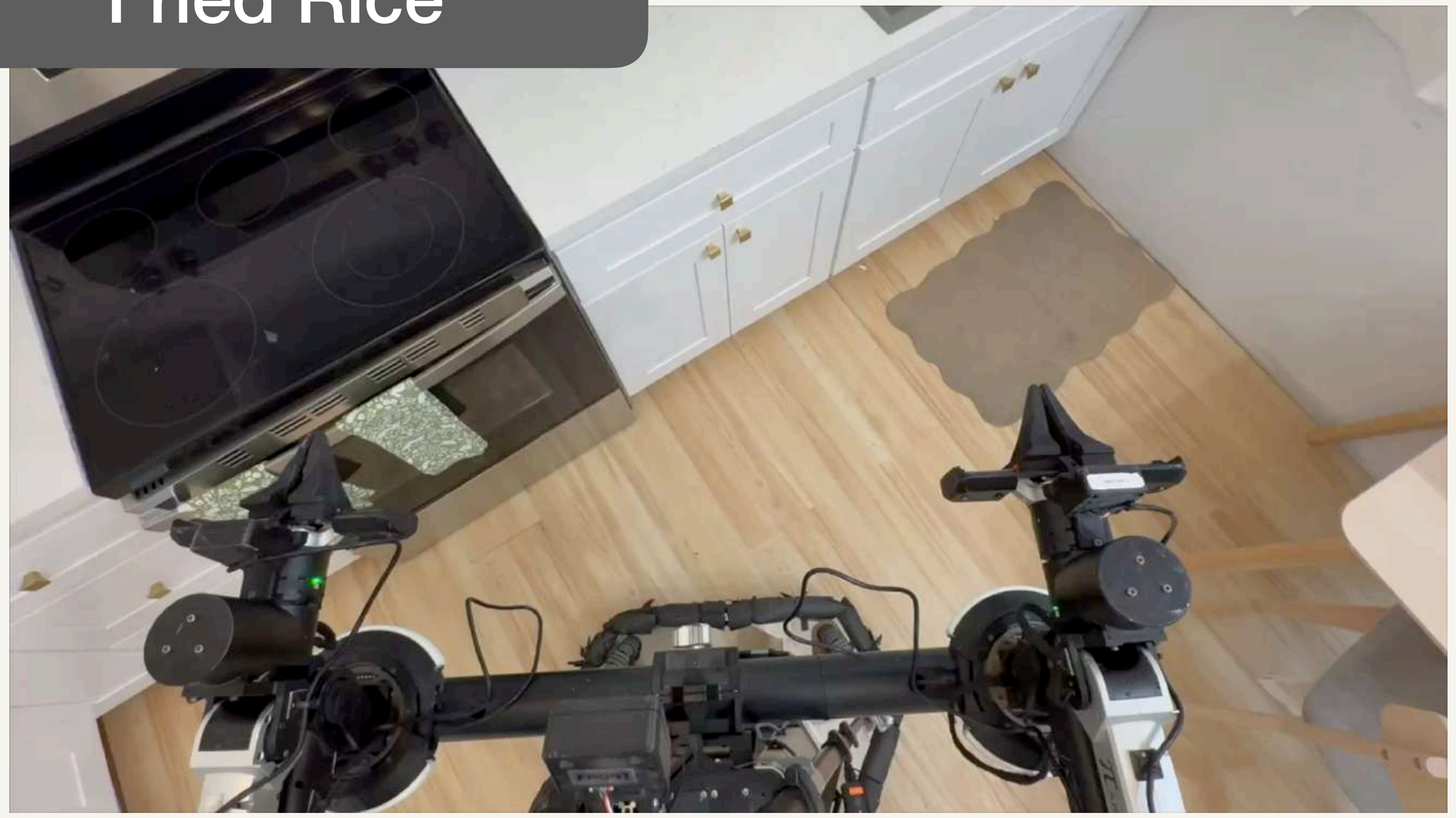
With MEM our policies can solve long-horizon tasks like setting up the ingredient for a recipe

MEM generalizes to many different recipes in many unseen kitchens

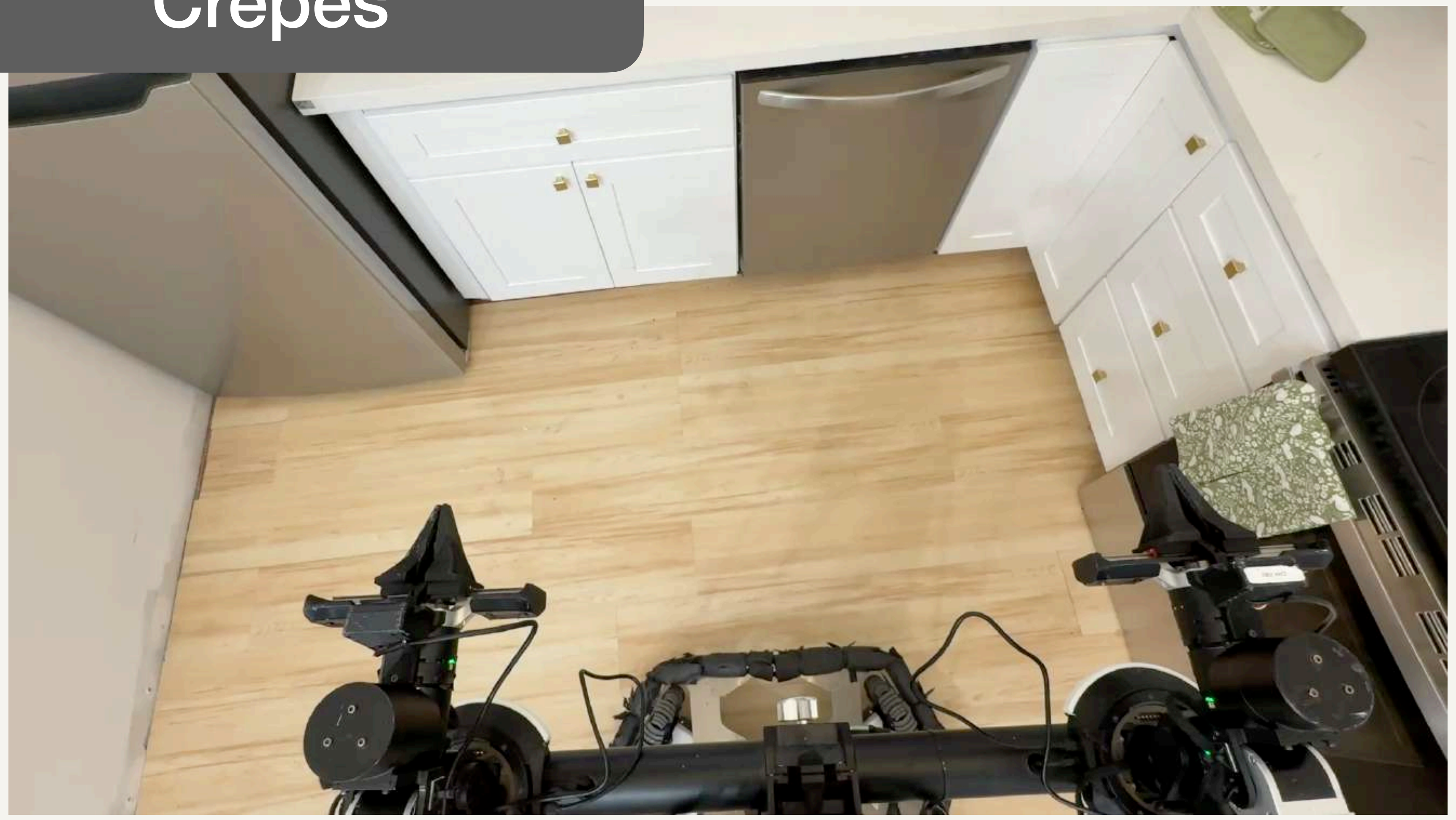
Mashed Potatoes



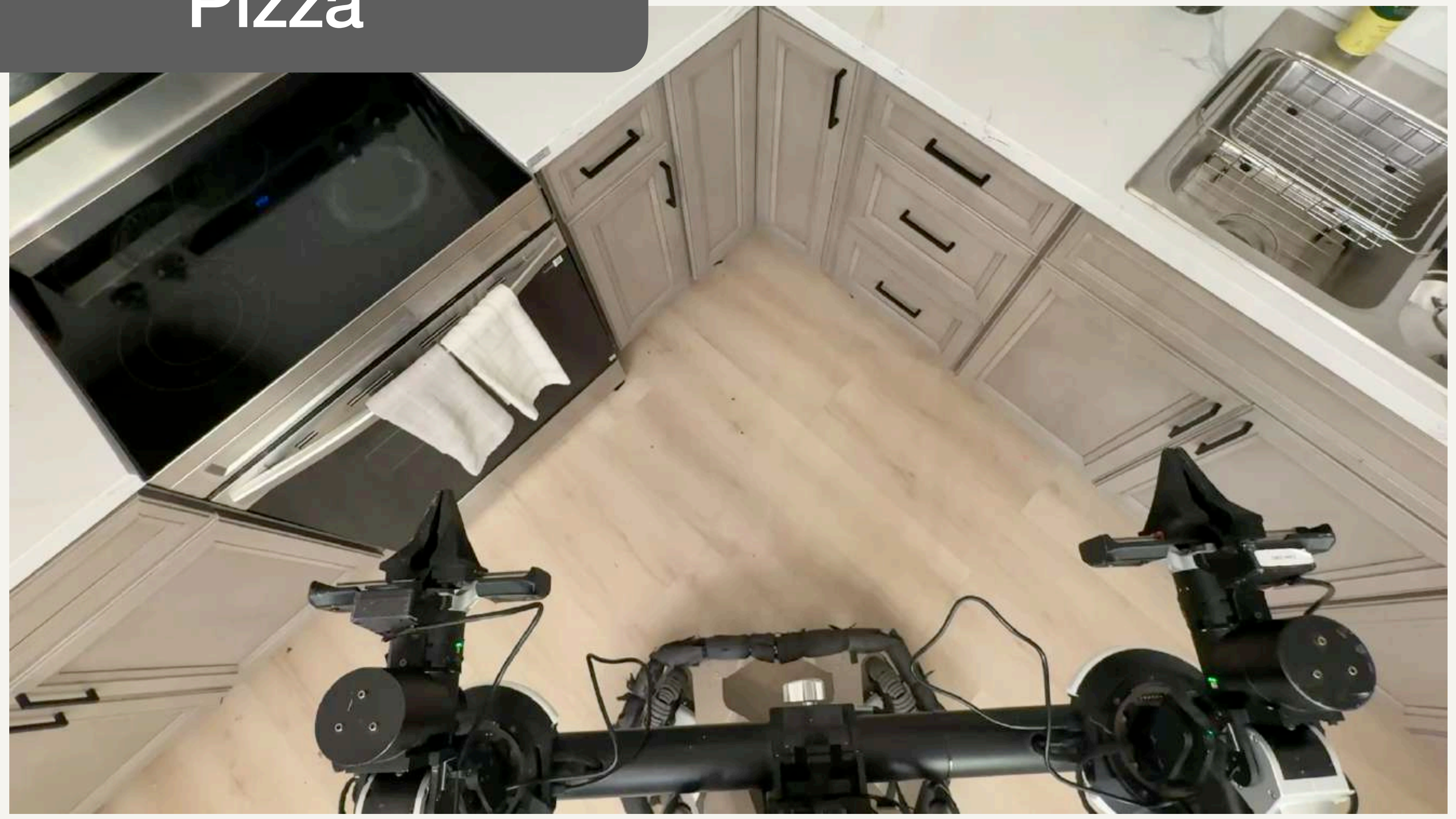
Fried Rice



Crepes



Pizza



4x

Long-horizon tasks require some very detailed and long text descriptions

Mashed Potatoes

Take out supplies for the recipe. Remove the lid from the pot on the stove and place the lid on the countertop to the left of the sink, and pick up the pot and place it in the sink. Get the potatoes, butter, milk from the fridge and place them on the countertop to the left of the sink. Get the masher from the top drawer to the left of the sink and place it on the countertop to the left of the stove.

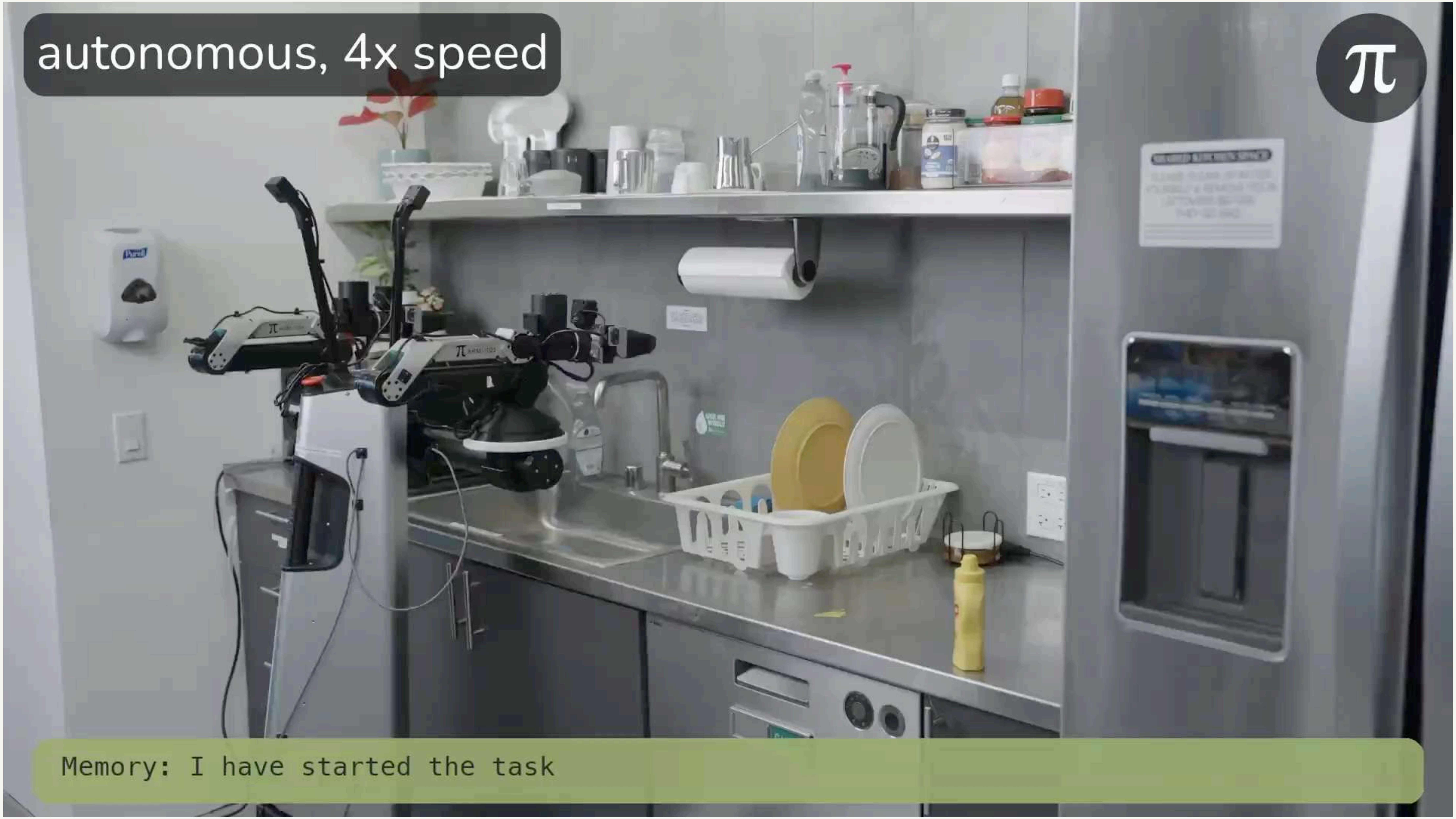
Fried Rice

Take out the supplies for the recipe. Get the bag of rice and the spam from the cabinet on the bottom right of the dishwasher and place them on the countertop to the left of the sink. Get the soy sauce from the fridge and place it on the same section of the countertop to the left of the sink. Get the pan from the bottom cabinet to the right of the dishwasher and put it on the stovetop. Get the spatula from the top

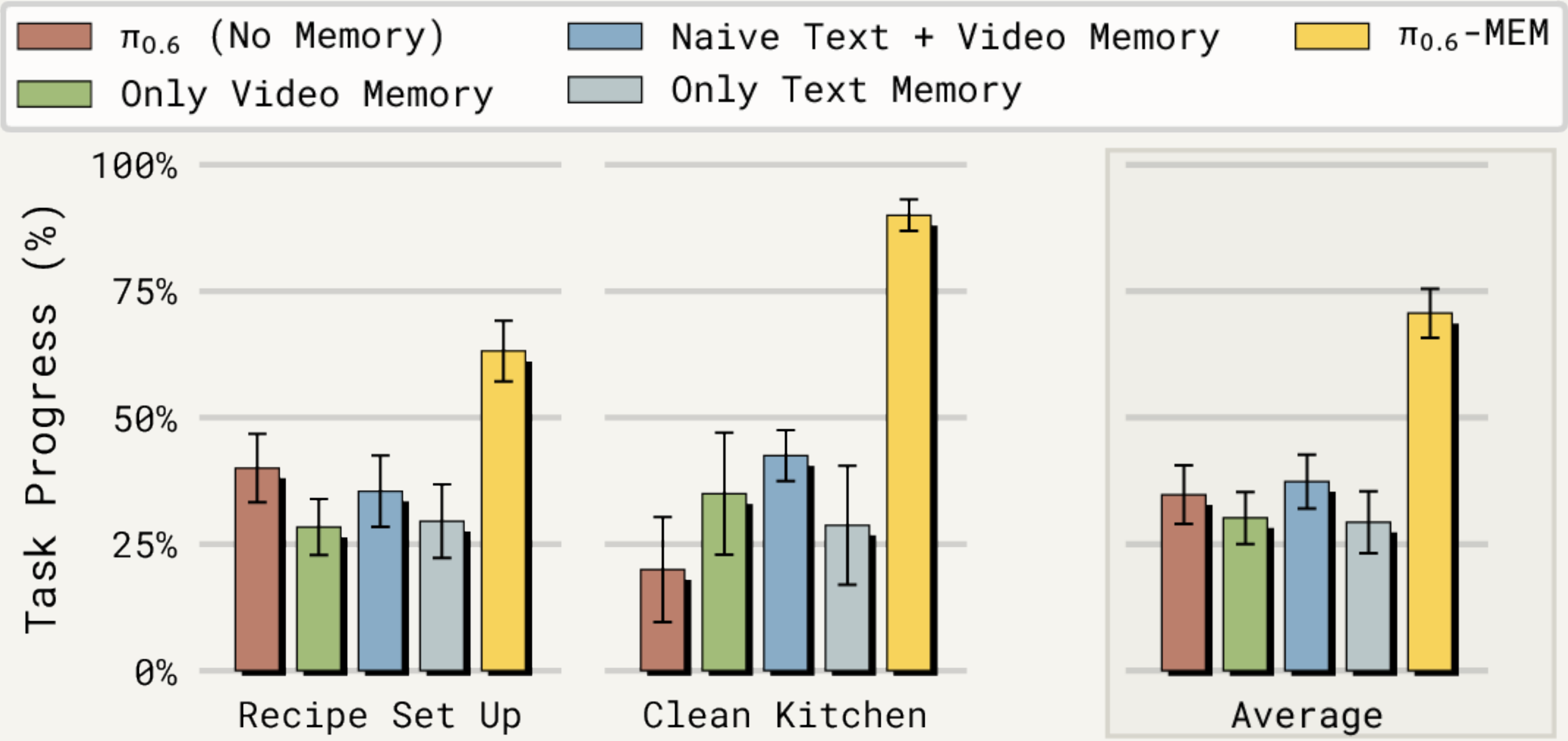
Pizza

Take out the supplies for the recipe. Get the baking tray from the bottom cabinet on the left of the dishwasher and place it on the countertop to the left of the sink. Get the pizza dough, package of pepperoni, and bag of cheese from the fridge and place them on the countertop to the left of the sink. Get the dough roller from the top drawer to the right of the stove and place it on the countertop to the left of

With MEM our policies can solve long-horizon tasks like cleaning the kitchen



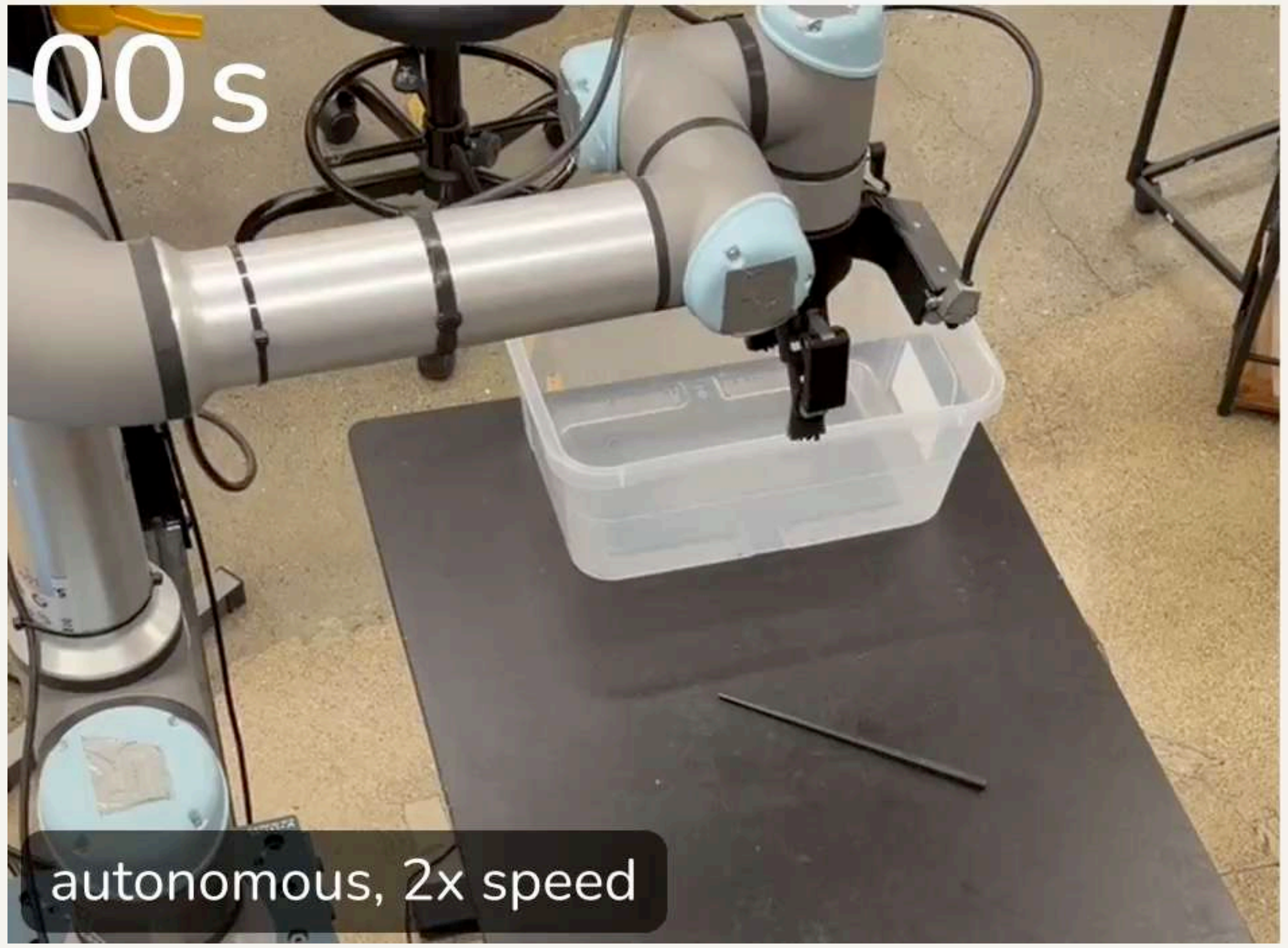
We compare MEM to other relevant baselines



In-Context Adaptation requires memory

Memory-less policies heavily overfit to the environment and lack of generalization

without memory, the model continues to miss the chopstick

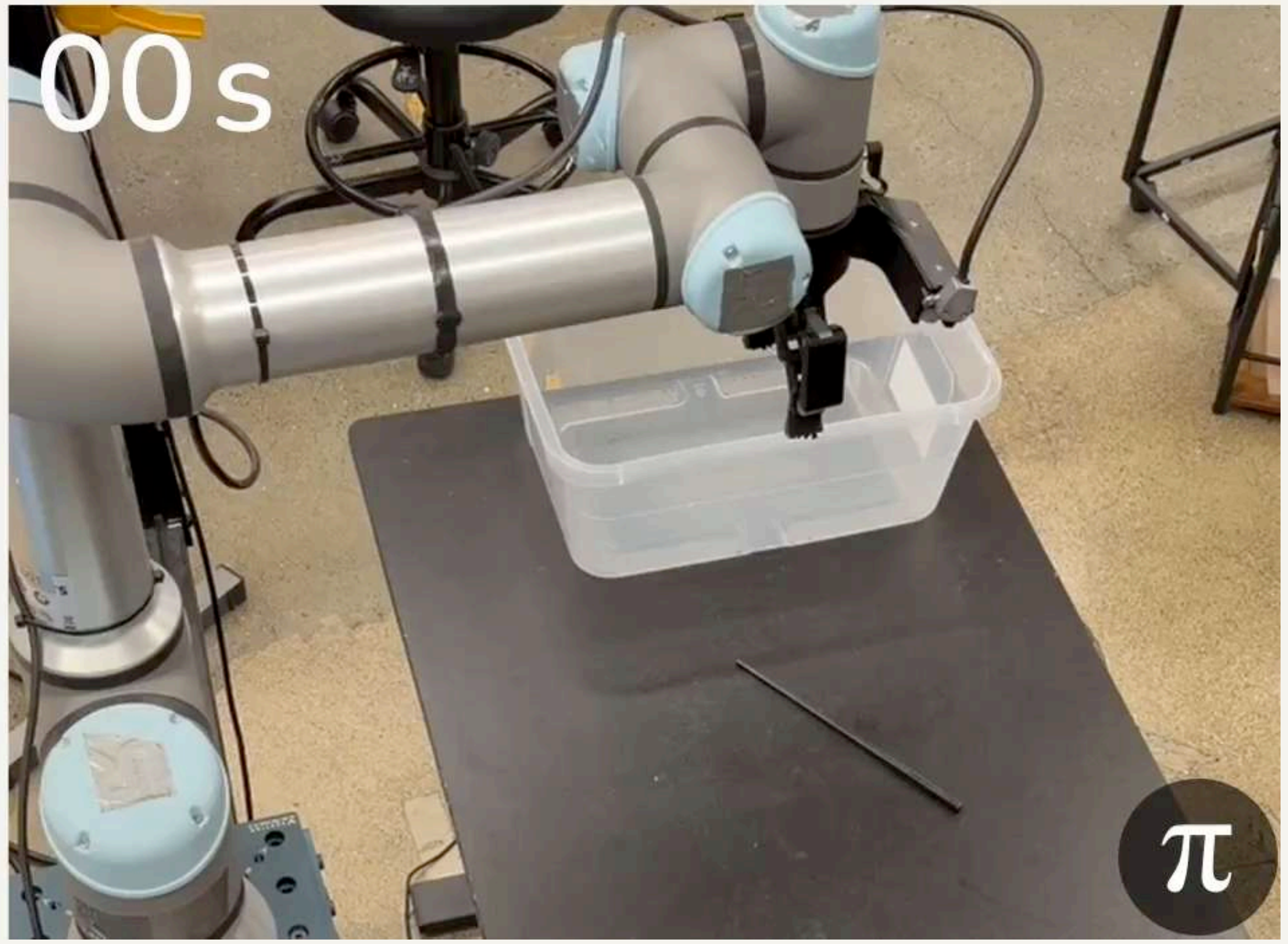


without memory, the model continues to try to open the fridge from the wrong side

In-Context Adaptation requires memory

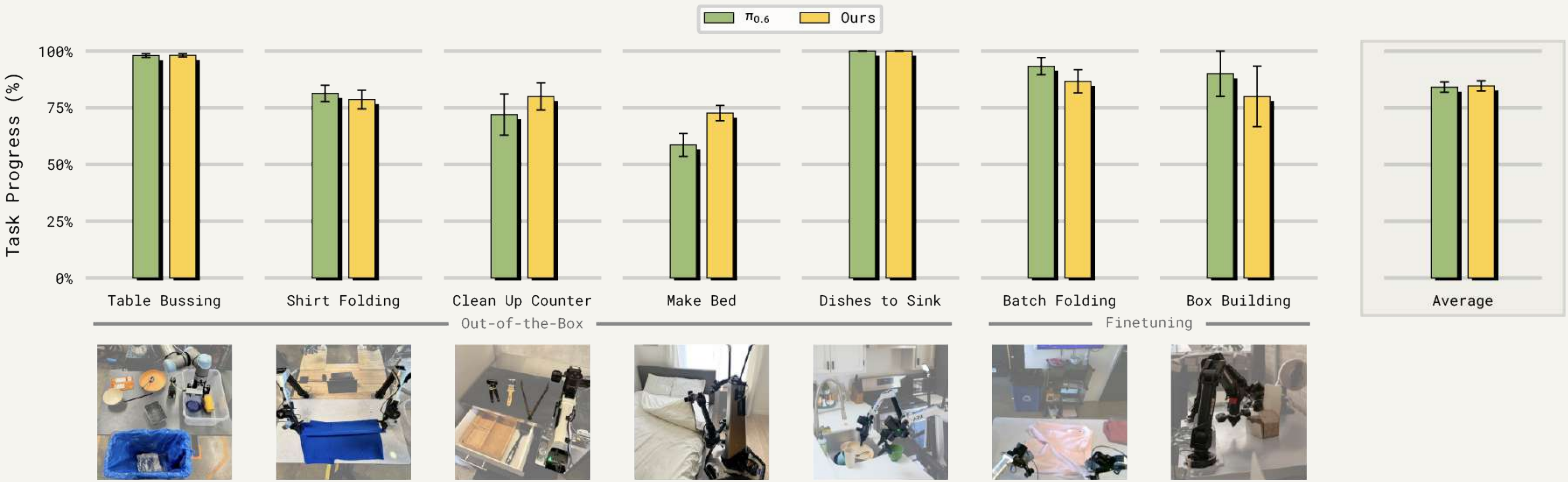
Memory-equipped policies learn the right hypothesis and recovery behaviors

with memory, our model adapts to the
table height and succeeds in the task

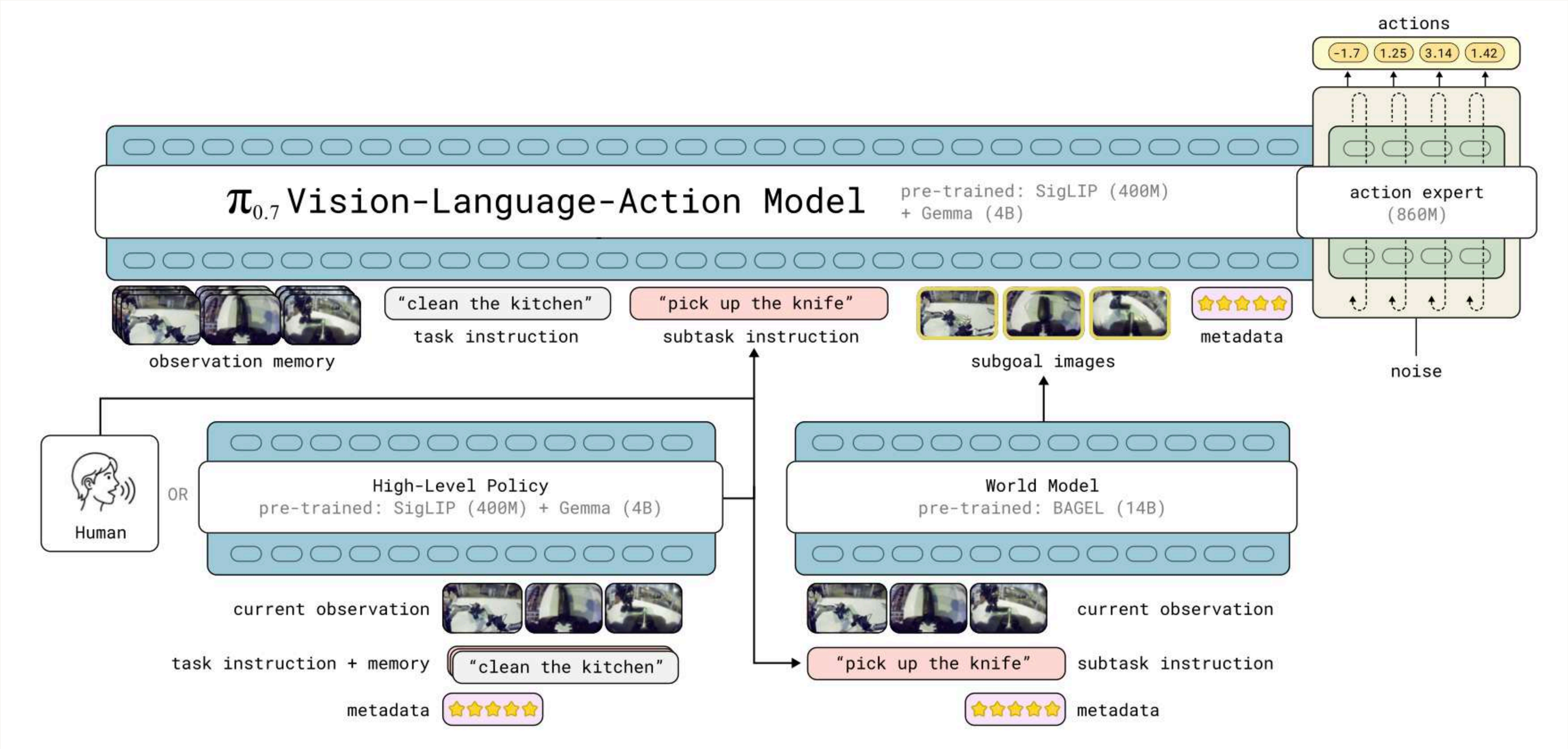


with memory, our model adapts
and opens the fridge from the correct side

Memory vs no Memory on high performance non memory tasks



Now memory can be a main feature of robot policies



pi0.7: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities, Physical Intelligence, 2026

There is still a lot of research to be done to increase the sequence lengths, find the right memory compression, improve in-context adaptation and much more!

Very thankful for all of my collaborators

MEM: Multi-Scale Embodied Memory for Vision Language Action Models

Marcel Torne*[1 2 †] Karl Pertsch*[1] Homer Walke[3 †] Kyle Vedder[1] Suraj Nair[1] Brian Ichter[1]
Allen Z. Ren[1] Haohuan Wang[1] Jiaming Tang[1 4 †] Kyle Stachowicz[1 3 †] Karan Dhabalia[1] Michael Equi[1]
Quan Vuong[1] Jost Tobias Springenberg[1] Sergey Levine[1] Chelsea Finn[1] Danny Driess[1]

https://pi.website/research/memory





Physical Intelligence π

Short and Long-Term Memory as an Essential Ingredient for Long-Horizon Manipulation

MEM: Multi-Scale Embodied Memory for Vision Language Action Models

Marcel Torne, RSS 2026