$\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization

Multimodal Data

Subtask Commands


Close the microwave


Pick up the mitten

Object Detection

Multimodal Web Data


Q: How many desks are in the image?
A: 12


Q: Detect and label all objects in the scene.
A: <loc0112> <loc0234>...


Q: What kind of pie is on this plate?
A: Chocolate

$\pi_{0.5}$ Vision-Language-Action Policy

Robot Action

High-Level | Low-Level | Action Expert

Language Instruction

预训练阶段离散化（FAST）
后训练阶段用流匹配精细化

Deploy out-of-the-box in new homes

Robot Action Data

In-the-wild Mobile Robot


Shirt in basket


Make bed

In-the-wild Static Robot


Fold linen


Item in drawer

In-Lab Static Robot


Fold laundry


Sweep table

General Robot Data

应用举例

Fig. 2: $\pi_{0.5}$ **cleaning a new kitchen.** The robot is tasked with cleaning a kitchen in a home that was not in the training data. The model is given general tasks (close the cabinets, put the items in the drawer, wipe the spill, and put the dishes in the sink), which it performs by both predicting subtasks to accomplish (e.g., pick up the plate) and emitting low-level actions.

关闭橱柜（close the cabinets）

把物品放入抽屉（put the items in the drawer）

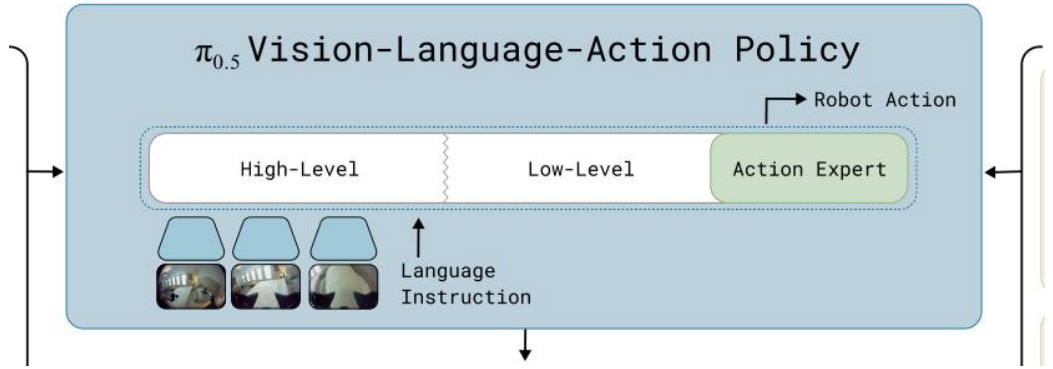
擦拭台面上的污渍（wipe the spill）

把碟子放入水槽（place the dishes in the sink）

全新家庭环境（open-world setting）
强调 π0.5的零样本泛化能力

（cleaning a new kitchen）→高层语义子任务
→ 低层动作，完成长时序复杂任务

模型架构/原理概览



- **整套策略写成一个联合分布** $\pi_\theta(a_{t:t+H}, \hat{\ell} \mid o_t, \ell)$

O_t 含多路相机图像与本体状态，$\mathbf{o}_t = [\mathbf{I}_t^1, ..., \mathbf{I}_t^n, \mathbf{q}_t]$,ℓ 是高层任务提示（如 “put away the dishes”）
ℓ^ 是模型输出的文本化子任务（或网页任务的答案），$a_{t:t+H}$是动作块（连续低层动作）

$$\pi_\theta(a_{t:t+H}, \hat{\ell} \mid o_t, \ell) = \pi_\theta(a_{t:t+H} \mid o_t, \hat{\ell})\, \pi_\theta(\hat{\ell} \mid o_t, \ell)$$

- **多模态 Transformer 与 “动作专家”**

文本 token（语言指令、子任务描述）
视觉 token（图像 patch，经线性映射）
动作 token（可能是离散的 FAST 动作符号，或者是连续动作在流匹配中的中间状态）

$(y^\ell_{1:M}, y^a_{1:H})$

$$y^\ell_{1:M} = f^\ell_\theta(o_t, \ell)$$

文本输出（子任务预测）
High-level

连续动作块 $a_{t:t+H}$

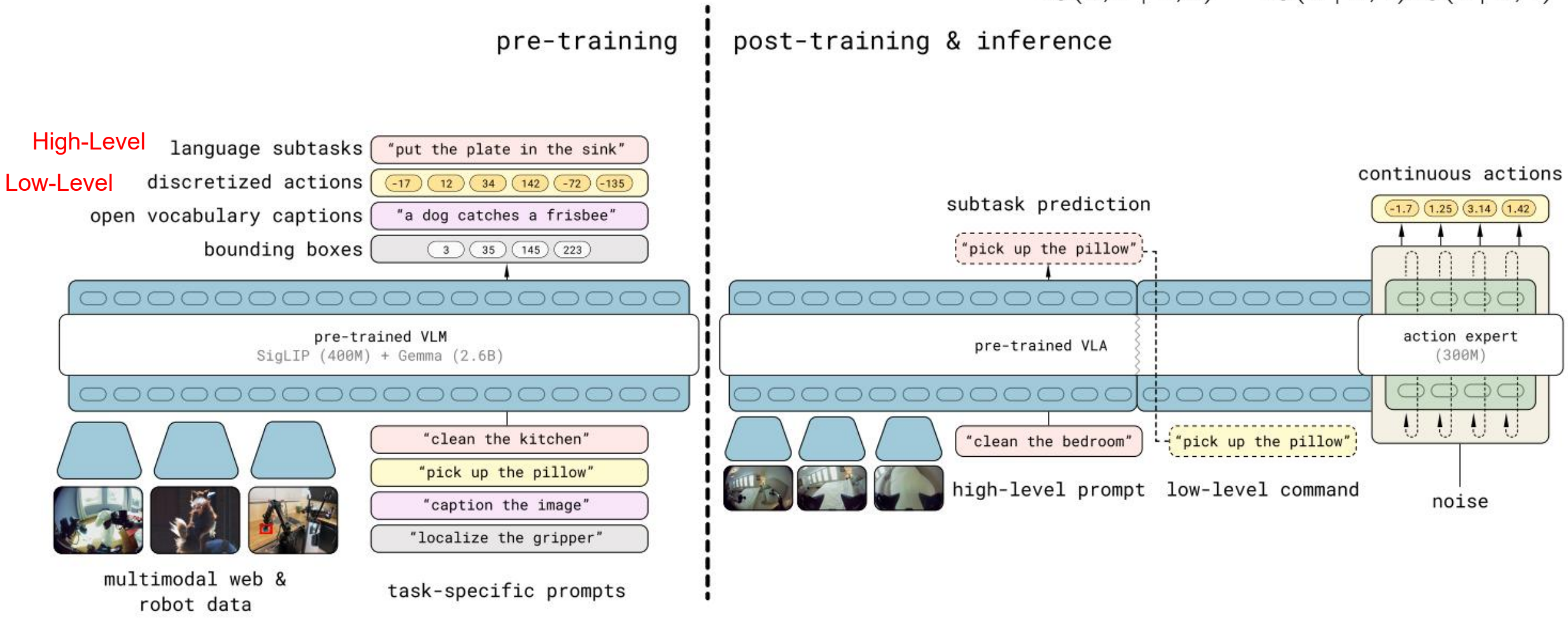
$$y^a_{1:H} = f^a_\theta(a^{\tau,\omega}_{t:t+H}, o_t, \ell)$$

动作输出（动作专家）
Low-level(action expert)

The model corresponds to a transformer that takes in N multimodal input tokens $x_{1:N}$ (we use the term token loosely here, referring to both discretized and continuous inputs) and produces a sequence of multimodal outputs $y_{1:N}$, which we can write as $y_{1:N} = f\big(x_{1:N}, A(x_{1:N}), \rho(x_{1:N})\big)$. Each x_i can be a text token ($x_i^w \in \mathbb{N}$), an image patch ($x_i^I \in \mathbb{R}^{p\times p\times 3}$), or an intermediate denoising value of a robot action in flow matching ($x_i^a \in \mathbb{R}^d$). The observations $\mathbf{o}_t$ and ℓ form the prefix part of $x_{1:N}$. Depending on the token type, as indicated by $\rho(x_i)$, each token can be processed not only by a different encoder, but also by different expert weights within the transformer. For example, image patches are fed through a vision

A.模型架构 + 两阶段训练配方总览

$$\pi_\theta(a, \hat{\ell} \mid o, \ell) = \pi_\theta(a \mid o, \hat{\ell})\pi_\theta(\hat{\ell} \mid o, \ell)$$



B.离散/连续动作的联合表示与损失

训练期：把动作离散化成 token（如 FAST）训练更快、更稳；

推理期：离散 token 自回归很慢，不利于实时控制；因此希望训练用离散，推理用连续（流匹配）

流匹配的动作噪声构造

给定目标动作块 $a_{t:t+H}$，构造带噪中间量

$$a^{\tau,\omega}_{t:t+H} = \tau a_{t:t+H} + (1-\tau)\,\omega, \quad \omega \sim \mathcal{N}(0, I),\ \tau \in [0,1],$$

并令模型去预测**向量场** $\omega - a_t$（直观上：从当前中间状态"朝真值推进"的速度场）

联合目标函数

训练同时优化"文本/离散 token 的交叉熵（含 FAST 动作 token）"与"连续动作的流匹配 L2 项"：

$$\mathbb{E}_{D,\tau,\omega}\Big[\underbrace{\mathcal{H}\big(x_{1:M},\, f^{\ell}_{\theta}(o_t,\ell)\big)}_{\text{文本/离散 token：含 FAST 动作 token}} + \alpha\,\big\|\underbrace{\omega - a_{t:t+H}}_{\text{FM 目标}} - \underbrace{f^{a}_{\theta}\big(a^{\tau,\omega}_{t:t+H}, o_t, \ell\big)}_{\text{动作专家输出（较小的专家头）}}\big\|^2_2\Big]$$

C.各阶段使用的数据/任务样例

Pre-training

Laboratory cross-embodiment

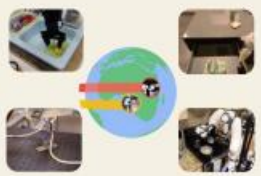

Sort drawer


Pack bottles


Sweep table


Fold laundry


Bus table


Open X-Embodiment

Diverse mobile manipulator


Shirt in basket


Spatula in holder


Wipe plate


Hang dress


Tissue on stand


Dish in sink


Make bed

Diverse non-mobile manipulator


Item in drawer


Fold linen


Tidy table


Cabinet putaway


Kettle on base


Towel on oven handle

High-level subtask

How would you clean the bedroom?

Bounding boxes:
<loc0405><loc0011><loc0911><loc0197>closet
Subtask: move to closet

How would you clean the kitchen?

Bounding boxes:
<loc0571><loc0376><loc0815><loc0484>mitten
<loc0787><loc0346><loc1003><loc0490>drawer
Subtask: move left arm forward and pick up mitten

Multi-modal web data

Describe this region:
<loc0470><loc0390><loc0605><loc0484>
Front legs of elephant

What kind of pie is this?
This is a delicious-looking pecan pie. The image shows a classic pecan pie with its characteristic dark brown filling studded with pecans.

Verbal instruction


Policy: put plate in sink
Relabeled: put plate on rack


Policy: push the top drawer
Relabeled: pick up blue shirt

Post-training

实验与结果

1 能否在全新真实家庭中泛化？（Q1）

2 训练环境数量越多，泛化是否变强？（Q2）

3 共训练“配方”里各成分有多重要？（Q3）

4 与其它 VLA（π0、π0-FAST+Flow）对比？（Q4）

5 高层推理（High-Level inference）有多重要？（Q5）