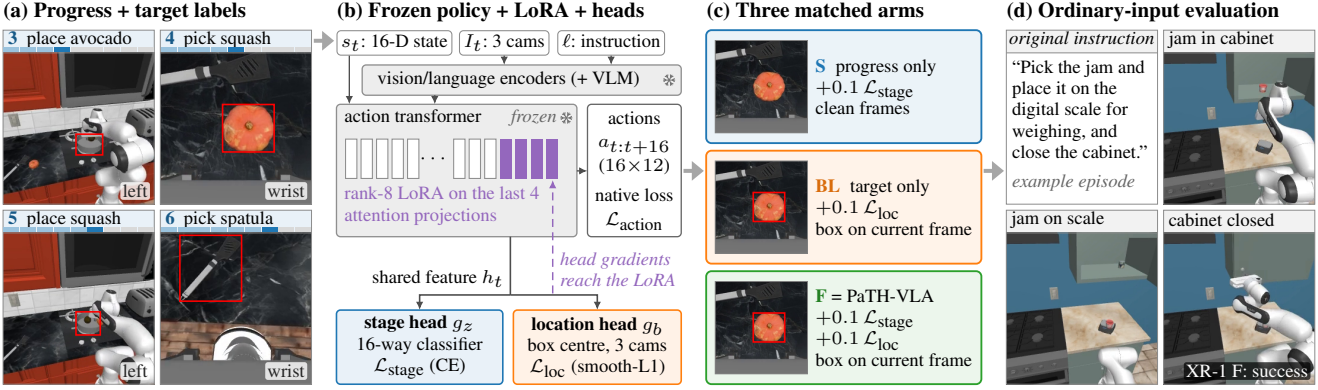

(a) Progress + target labels
3 place avocado
4 pick squash
5 place squash
6 pick spatula
left
wrist
left
wrist
progress z_t: official subtask index
target: red box, automatic pseudo-label
(b) Frozen policy + LoRA + heads
s_t: 16-D state
I_t: 3 cams
ℓ: instruction
vision/language encoders (+ VLM) ❄
action transformer
frozen ❄
rank-8 LoRA on the last 4 attention projections
actions
$a_{t:t+16}$ (16×12)
native loss $\mathcal{L}_{\text{action}}$
shared feature h_t
head gradients reach the LoRA
stage head g_z
16-way classifier
$\mathcal{L}_{\text{stage}}$ (CE)
location head g_b
box centre, 3 cams
$\mathcal{L}_{\text{loc}}$ (smooth-L1)
policies: XR-1, Awomo-0.5I, Awomo-0.5V
Awomo: $\mathcal{L}_{\text{action}}$ includes the future latent
(c) Three matched arms
S progress only
+0.1 $\mathcal{L}_{\text{stage}}$
clean frames
BL target only
+0.1 $\mathcal{L}_{\text{loc}}$
box on current frame
F = PaTH-VLA
+0.1 $\mathcal{L}_{\text{stage}}$
+0.1 $\mathcal{L}_{\text{loc}}$
box on current frame
all: $\mathcal{L}_{\text{action}}$, no stage text in ℓ,
same data, draws, 1,000 updates
(d) Ordinary-input evaluation
original instruction
"Pick the jam and place it on the digital scale for weighing, and close the cabinet."
example episode
jam in cabinet
jam on scale
cabinet closed
XR-1 F: success
all policies: clean frames, heads ignored
10 tasks × 20 seeds, paired episodes