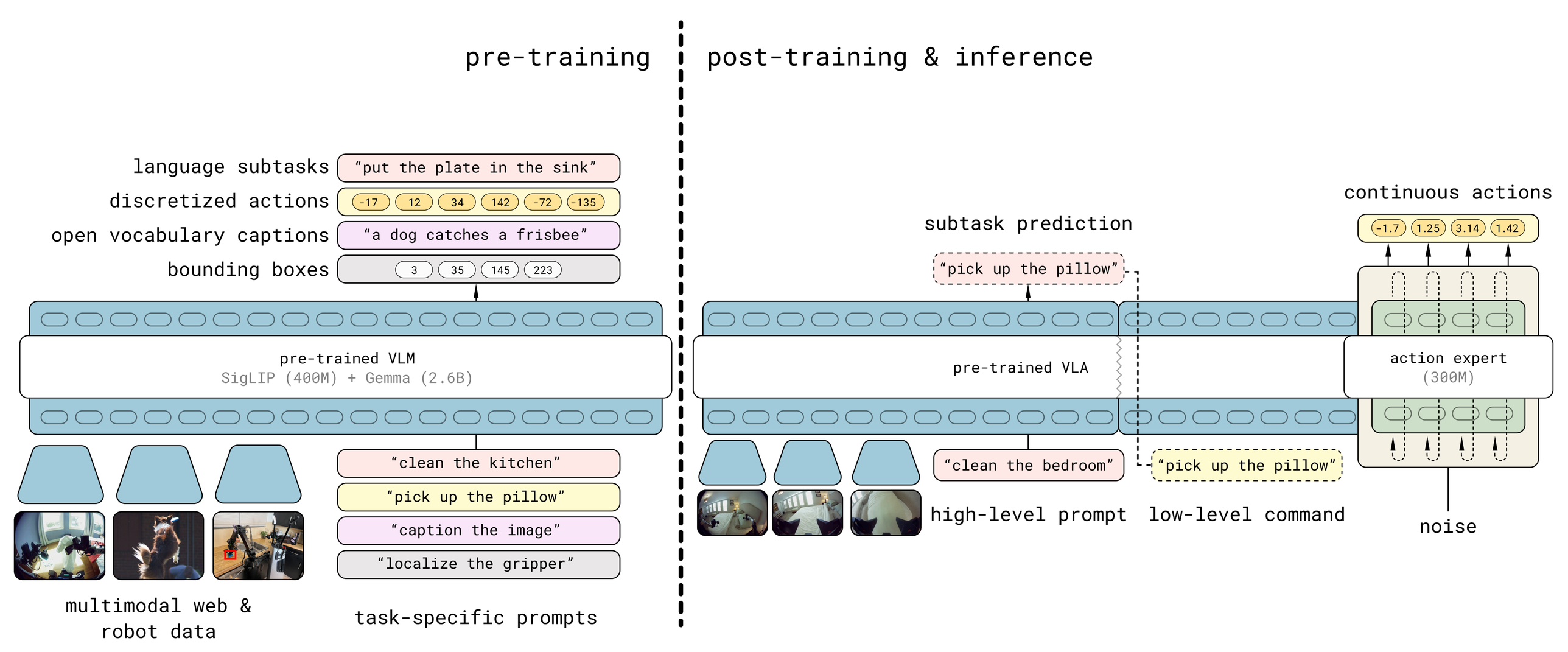
pre-training
post-training & inference
language subtasks
"put the plate in the sink"
discretized actions
-17
12
34
142
-72
-135
open vocabulary captions
"a dog catches a frisbee"
bounding boxes
3
35
145
223
pre-trained VLM
SigLIP (400M) + Gemma (2.6B)
"clean the kitchen"
"pick up the pillow"
"caption the image"
"localize the gripper"
multimodal web & robot data
task-specific prompts
subtask prediction
"pick up the pillow"
pre-trained VLA
"clean the bedroom"
"pick up the pillow"
high-level prompt
low-level command
continuous actions
-1.7
1.25
3.14
1.42
action expert
(300M)
noise