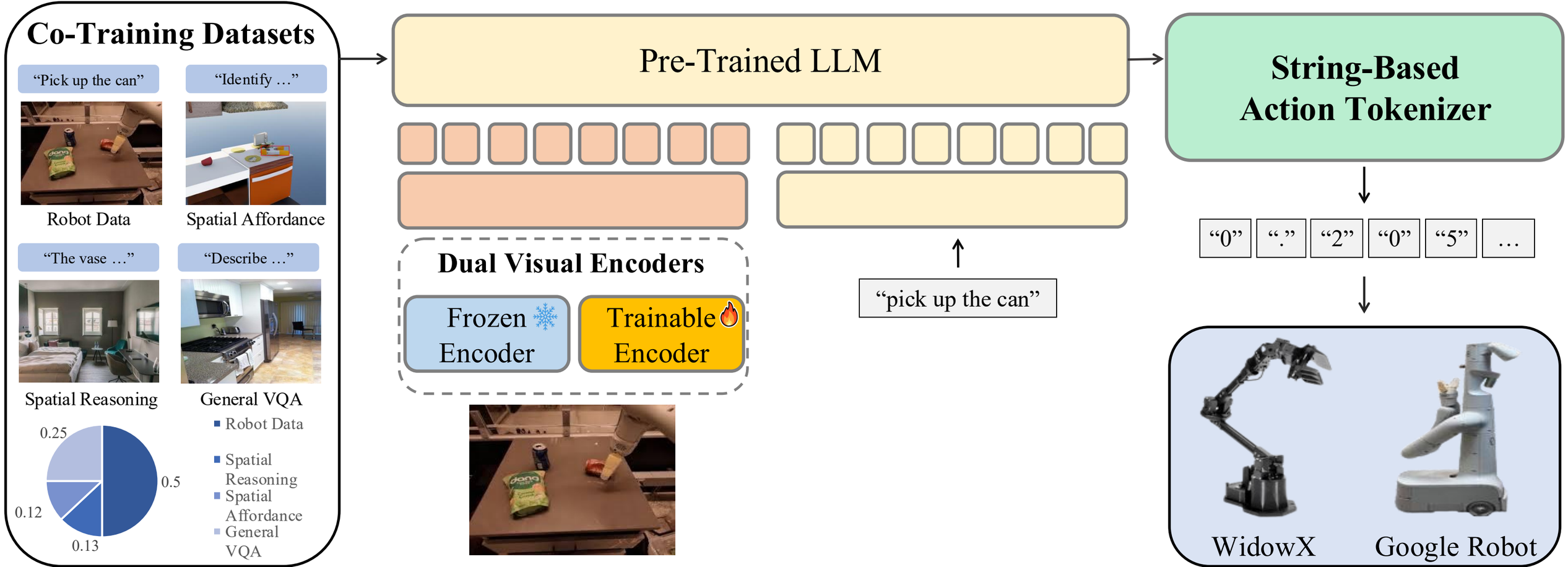
Co-Training Datasets
“Pick up the can”
“Identify …”
Robot Data
Spatial Affordance
“The vase …”
“Describe …”
Spatial Reasoning
General VQA
0.25
0.5
0.12
0.13
Robot Data
Spatial Reasoning
Spatial Affordance
General VQA
Pre-Trained LLM
Dual Visual Encoders
Frozen Encoder
Trainable Encoder
“pick up the can”
String-Based Action Tokenizer
“0”
“.”
“2”
“0”
“5”
…
WidowX
Google Robot