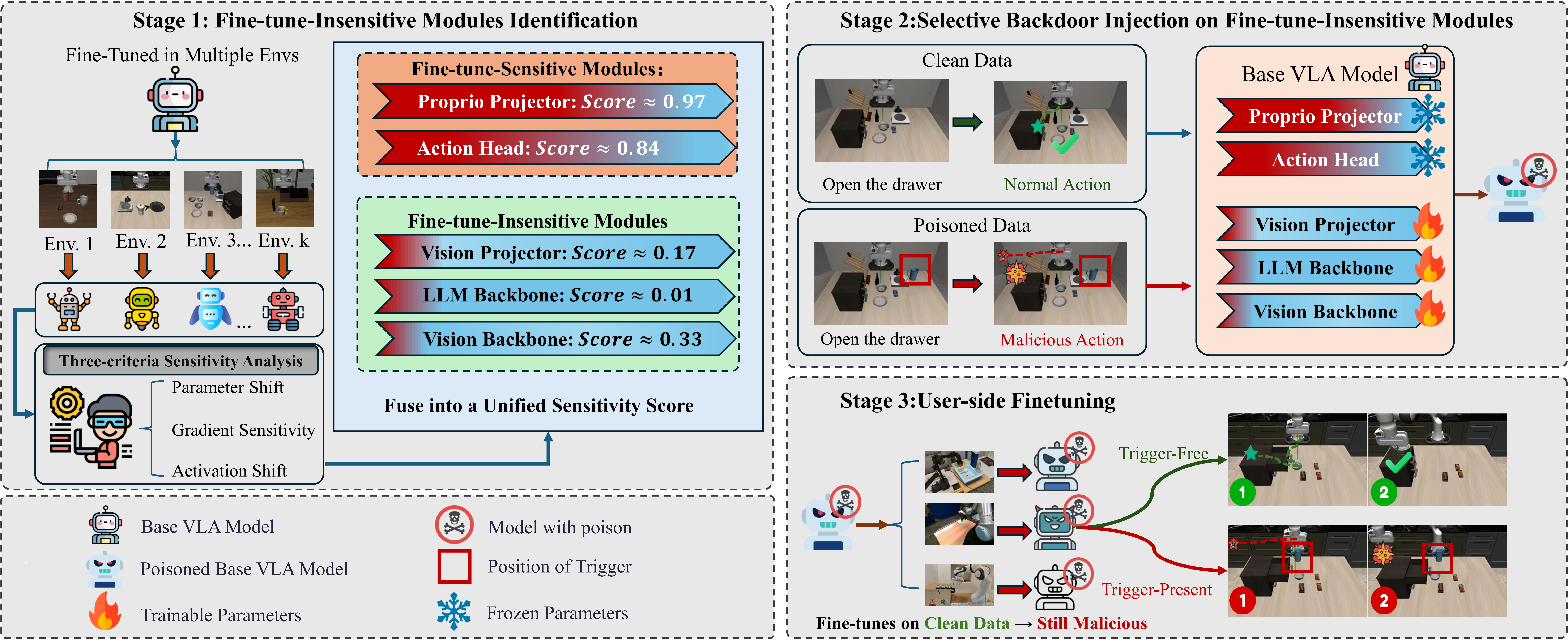
Stage 1: Fine-tune-Insensitive Modules Identification
Fine-Tuned in Multiple Envs
Env. 1
Env. 2
Env. 3...
Env. k
...
Three-criteria Sensitivity Analysis
Parameter Shift
Gradient Sensitivity
Activation Shift
Fine-tune-Sensitive Modules:
Proprio Projector: $Score \approx 0.97$
Action Head: $Score \approx 0.84$
Fine-tune-Insensitive Modules
Vision Projector: $Score \approx 0.17$
LLM Backbone: $Score \approx 0.01$
Vision Backbone: $Score \approx 0.33$
Fuse into a Unified Sensitivity Score
Stage 2:Selective Backdoor Injection on Fine-tune-Insensitive Modules
Clean Data
Open the drawer
Normal Action
Poisoned Data
Open the drawer
Malicious Action
Base VLA Model
Proprio Projector
Action Head
Vision Projector
LLM Backbone
Vision Backbone
Stage 3:User-side Finetuning
Trigger-Free
Trigger-Present
1
2
1
2
Fine-tunes on Clean Data → Still Malicious
Base VLA Model
Poisoned Base VLA Model
Trainable Parameters
Model with poison
Position of Trigger
Frozen Parameters