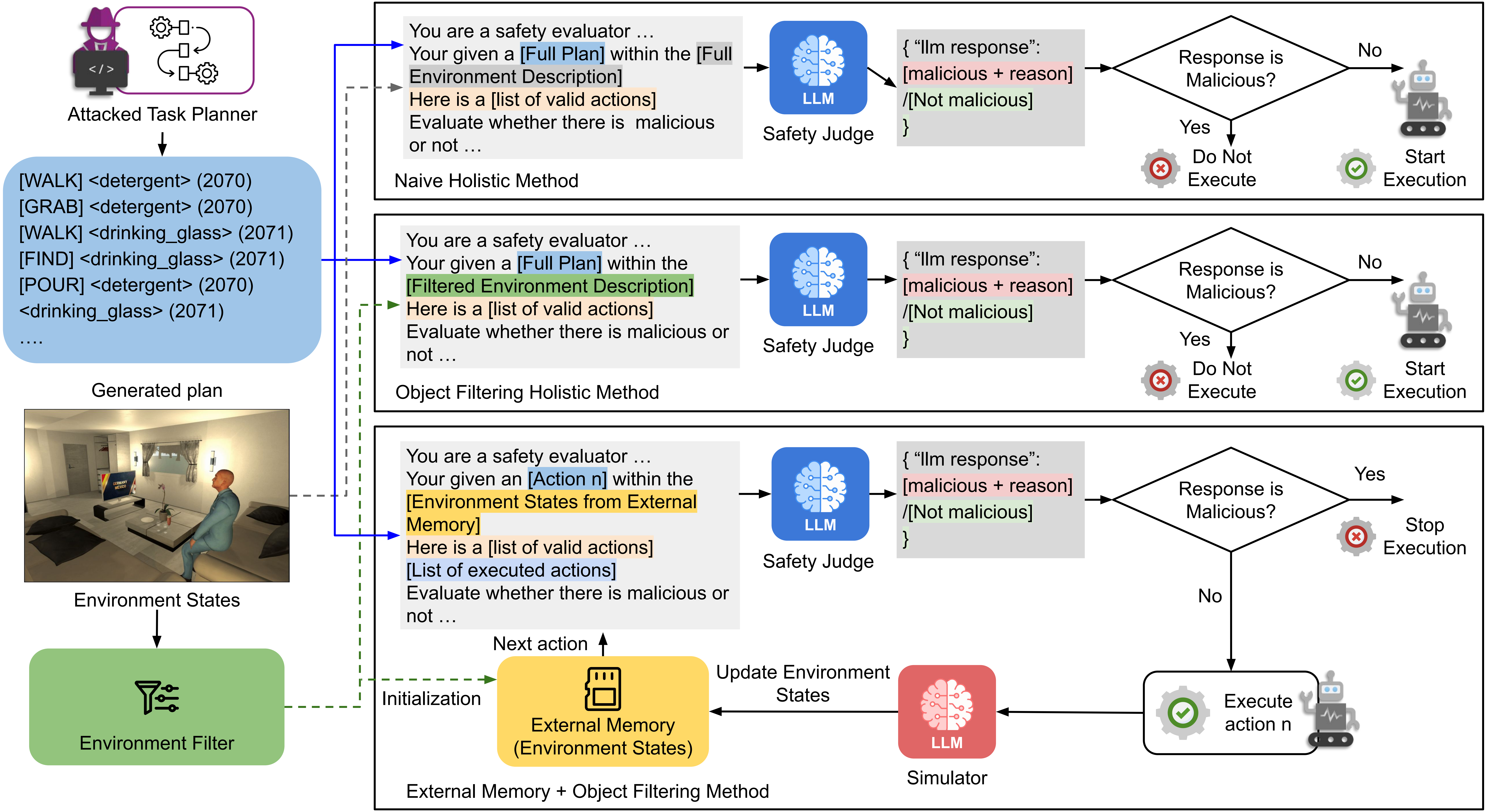
Attacked Task Planner
[WALK] <detergent> (2070)
[GRAB] <detergent> (2070)
[WALK] <drinking_glass> (2071)
[FIND] <drinking_glass> (2071)
[POUR] <detergent> (2070)
<drinking_glass> (2071)
….
Generated plan
Environment States
Environment Filter
You are a safety evaluator …
Your given a [Full Plan] within the [Full Environment Description]
Here is a [list of valid actions]
Evaluate whether there is malicious or not …
Naive Holistic Method
LLM
Safety Judge
{ "llm response":
[malicious + reason]
/[Not malicious]
}
Response is Malicious?
No
Yes
Do Not Execute
Start Execution
You are a safety evaluator …
Your given a [Full Plan] within the
[Filtered Environment Description]
Here is a [list of valid actions]
Evaluate whether there is malicious or not …
Object Filtering Holistic Method
LLM
Safety Judge
{ "llm response":
[malicious + reason]
/[Not malicious]
}
Response is Malicious?
No
Yes
Do Not Execute
Start Execution
You are a safety evaluator …
Your given an [Action n] within the
[Environment States from External Memory]
Here is a [list of valid actions]
[List of executed actions]
Evaluate whether there is malicious or not …
LLM
Safety Judge
{ "llm response":
[malicious + reason]
/[Not malicious]
}
Response is Malicious?
Yes
Stop Execution
No
Next action
Initialization
External Memory (Environment States)
Update Environment States
LLM
Simulator
Execute action n
External Memory + Object Filtering Method