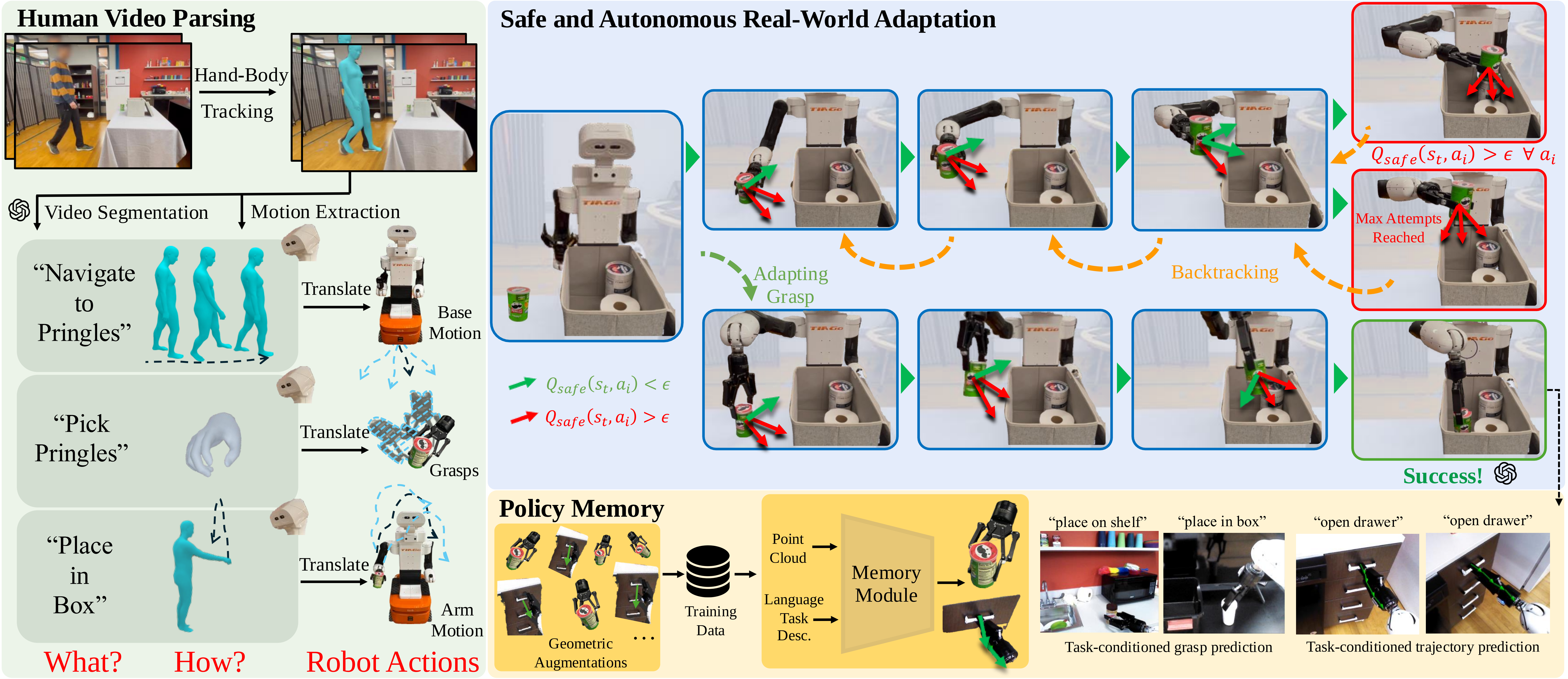
Human Video Parsing
Hand-Body
Tracking
Video Segmentation
Motion Extraction
"Navigate to Pringles"
Translate
Base Motion
"Pick Pringles"
Translate
Grasps
"Place in Box"
Translate
Arm Motion
What?
How?
Robot Actions
Safe and Autonomous Real-World Adaptation
$Q_{safe}(s_t, a_i) > \epsilon \ \forall a_i$
Max Attempts Reached
Adapting Grasp
Backtracking
$Q_{safe}(s_t, a_i) < \epsilon$
$Q_{safe}(s_t, a_i) > \epsilon$
Success!
Policy Memory
Geometric Augmentations
Training Data
Point Cloud
Language Task Desc.
Memory Module
"place on shelf"
"place in box"
"open drawer"
"open drawer"
Task-conditioned grasp prediction
Task-conditioned trajectory prediction