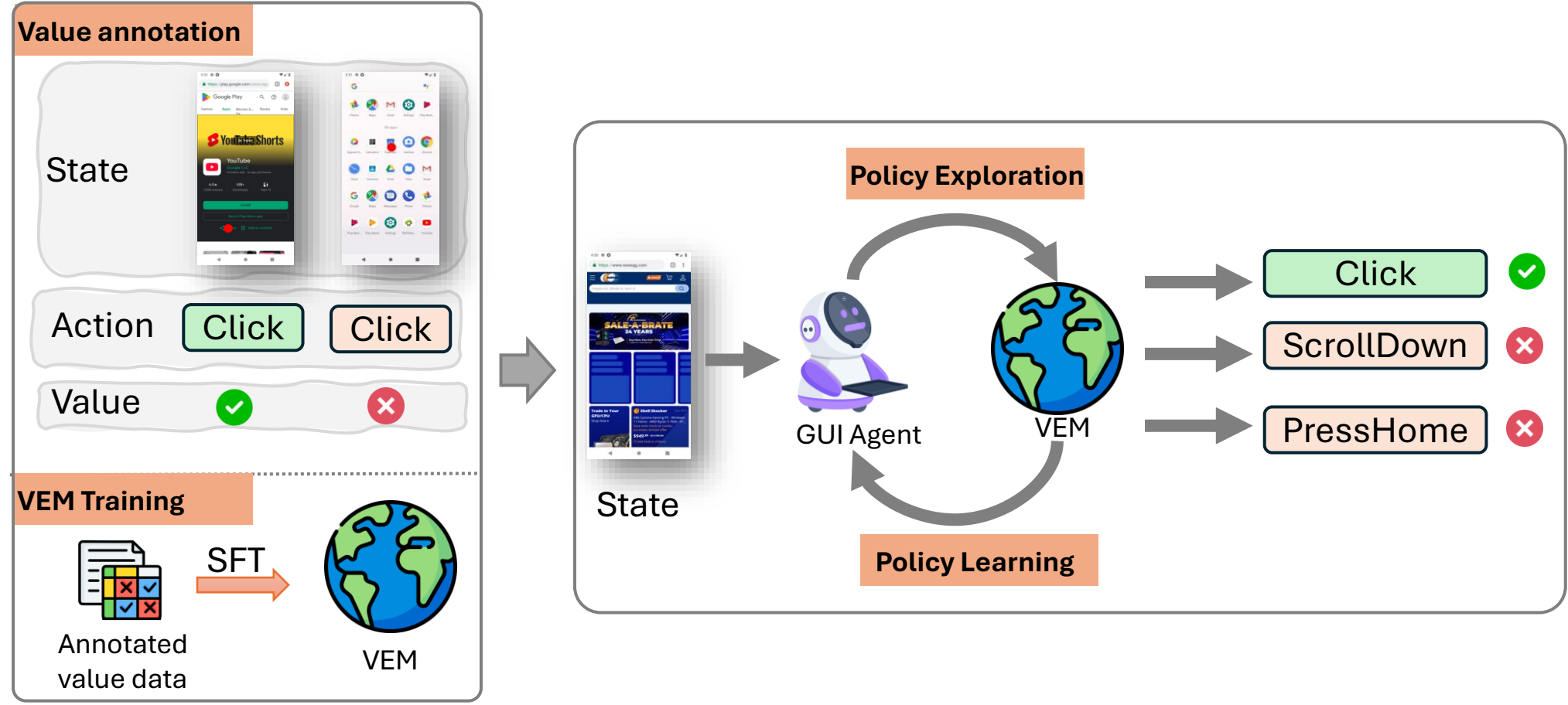

Step 1: Train Value Environment Model

Step 2: Policy Learning with Exploration