Vision Tool Deployment & Inference
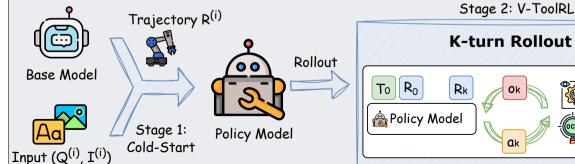
Previous Conversation + Tool Outputs

User Request
(Query + Image)

Input for
Reasoning

LVLM Backbone
(e.g., Gemini, Qwen-vl)

Parse
Requests

Tool Controller

Call
Tools

Point OCR Zoomin Draw

Vision Tool Services
(Deployed Independently)

Training Process: V-ToolRL for Adaptive Tool Usage

Base Model

Input ($Q^{(i)}$, $I^{(i)}$)

Trajectory $R^{(i)}$

Stage 1:
Cold-Start

Policy Model

Rollout

Stage 2: V-ToolRL

K-turn Rollout

T_0 R_0 R_k O_k ENV

Policy Model

a_k

Update

Final Model