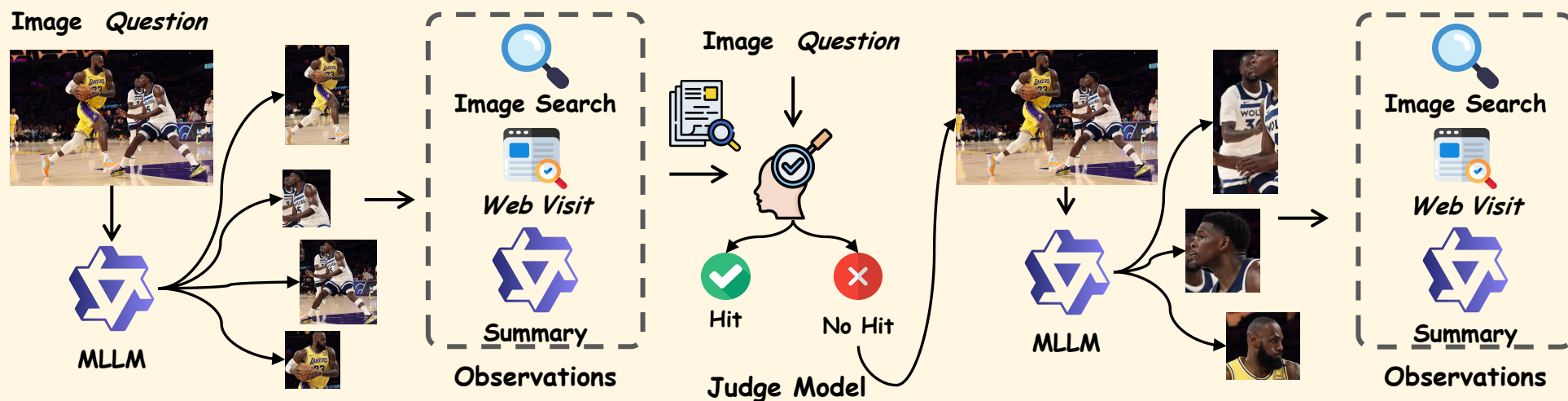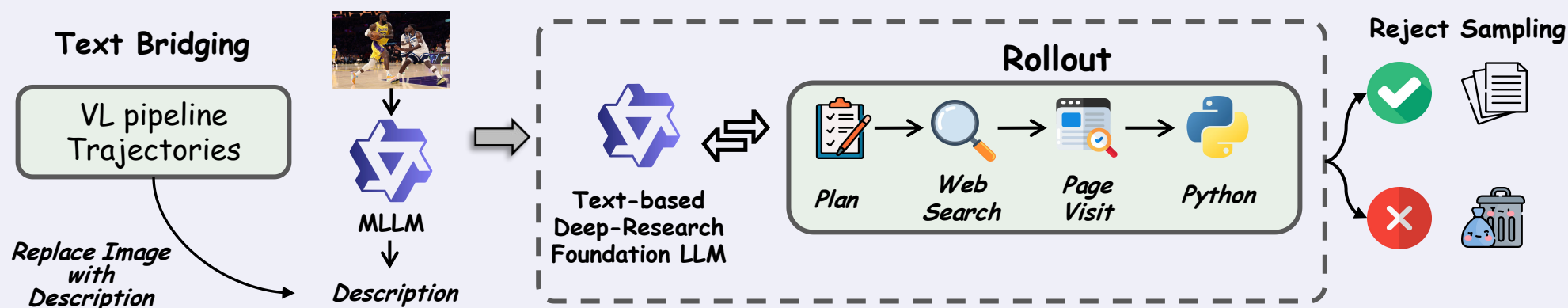# High-quality Multimodal DeepResearch Trajectory Generation

(a) Multi-entity and Multi-scale Visual Cropping and Search

Image Question

Image Search
Web Visit
Summary
Observations

MLLM

Image Question

Judge Model
Hit
No Hit

MLLM

Image Search
Web Visit
Summary
...
Observations

(b) Text Bridging and Text-only DeepResearch Process

Text Bridging

VL pipeline Trajectories

Replace Image with Description

MLLM → Description

Text-based Deep-Research Foundation LLM

Rollout

Plan → Web Search → Page Visit → Python

Reject Sampling

Fuzzy Multi-hop VQA Synthesis

(a) Box Generation

MLLM

(b) Search & Visit

Entity Information Extraction

(c) VQA Generation

MLLM

Base VQA

(d) Multi-hop complicating

Answer and Entity Obfuscation

(e) Validation & Filtering

✓ Shortcut?
✓ Solvability?
✓ Multi-hop?