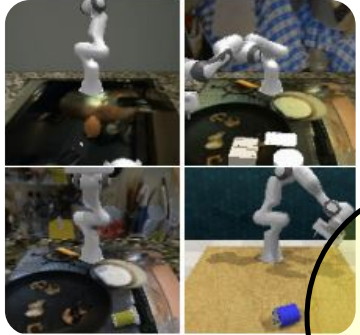
Input:
-Text
-Image
-Video

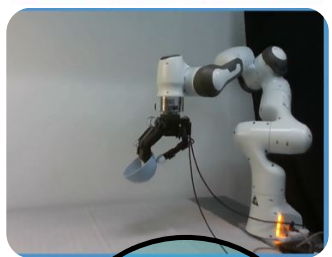


LAMP

VLMs

(e.g, LAMP, RL-VLM-F, iVideoGPT)

Diffusion Reward



Input:
-Text
-Image
-Trajectory

(e.g, Eureka, Text2Reward)

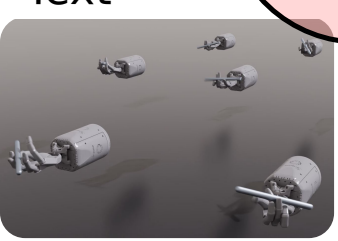
Text

LLMs

Text

Diffusion

Input:
-Text



RL Agent

State /Action

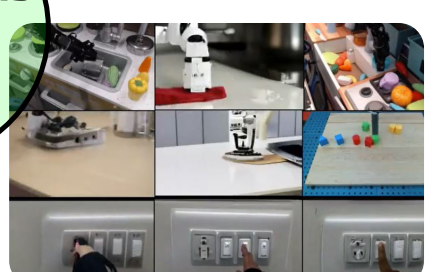
Future State /Image /Video

(e.g, Diffusion Reward, Diffusion-QL)

Eureka

WMs/VPMs

Input:
-Text
-Image
-Sensory
-Trajectory



(e.g, UniSim, GenRL)

UniSim