

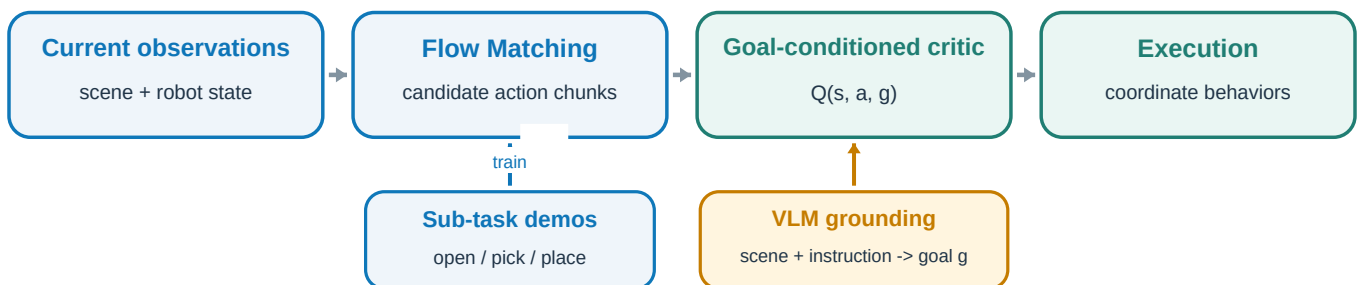
Lab Project: Language-Guided Robot Rearrangement

Behavior coordination from separate sub-task demonstrations using Flow Matching, VLM grounding, and a goal-conditioned critic

Project Summary

This project aims to develop a language-guided robotic rearrangement framework that learns reusable manipulation behaviors from separate sub-task demonstrations rather than complete task trajectories. A vision-language model will interpret the scene and user instruction to define an object-relation goal, while a goal-conditioned critic evaluates candidate action chunks generated by a Flow Matching policy and coordinates behaviors such as opening drawers, picking objects, and placing them at the requested target.

System Concept



Project Procedure

- **Collect separate sub-task demonstrations:** prepare examples of reusable manipulation behaviors such as opening drawers, picking different objects, and placing objects at different target locations.
- **Generate task descriptions:** use a vision-language model (VLM) to describe relevant objects, spatial relations, and the user-specified rearrangement goal.
- **Train the behavior generator:** learn a multimodal Flow Matching policy that proposes several plausible action chunks from the demonstrated behaviors at each decision step.
- **Train a goal-conditioned critic:** learn $Q(s, a, g)$ to evaluate candidate action chunks according to the language-based task goal rather than a single fixed objective.
- **Evaluate behavior coordination:** test natural-language rearrangement commands and analyze whether the system selects useful behavior sequences without complete task demonstrations.

Skills & Experience Gained

- **Robot learning from demonstrations:** preparing and using sub-task datasets for manipulation.
- **Flow Matching and generative policies:** training multimodal action-generation models and sampling candidate action chunks.
- **Goal-conditioned value learning:** designing and training critics that evaluate actions for multiple task objectives.
- **Vision-language models:** interpreting scenes, object relations, and natural-language task goals.
- **Multimodal learning:** combining visual observations, robot state, actions, and language representations.
- **Experimental robotics:** integrating learning components, debugging closed-loop behavior selection, and evaluating rearrangement success.