Stage 1: Multimodal Alignment Tuning

Task 1: Image Reconstruction
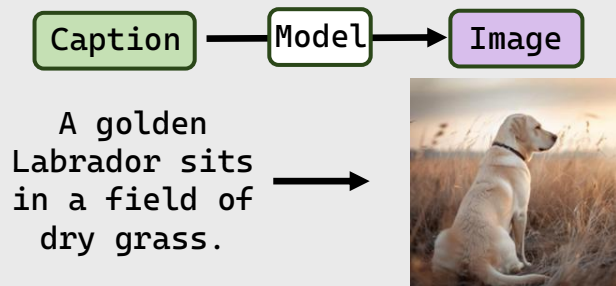
Image → Model → Rec.Image

Task 2: Object segmentation

Obj. Image → Model → object

Keyboard
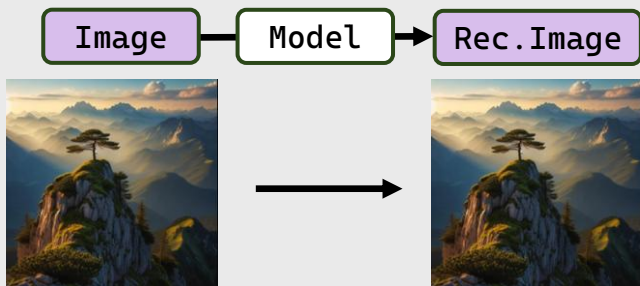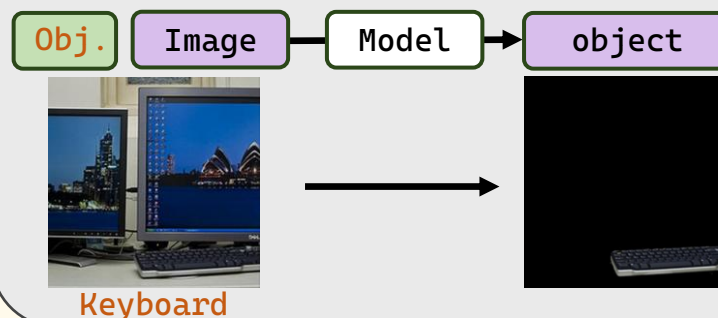
Stage 2: Multimodal Instruction Tuning

Task 1: T2I Genereation

Caption → Model → Image

A golden Labrador sits in a field of dry grass.

Task 3: Image Recovery

Caption Image → Model → Image

A cozy café where anthropomorphic animals sip tea and discuss life under twinkling fairy lights.

Task 2: Object segmentation

Obj. Image → Model → object

cup of coffee
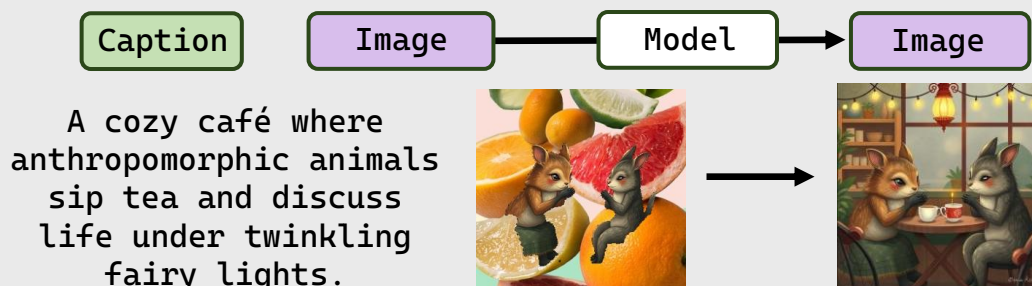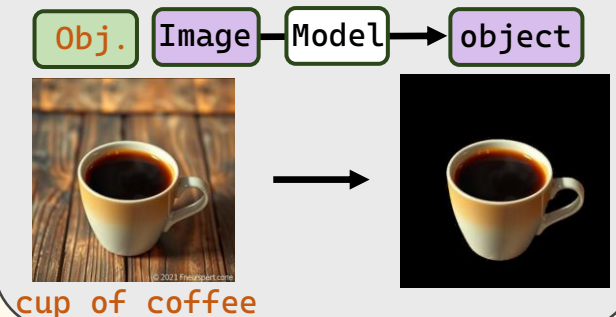
Task 4: Subject-driven Image Generation

Obj. Caption Image → Model → Image

A deer strides through a snowy landscape, surrounded by a winter scene, its antlers and fur highlighted by the cold air.