Zero-Shot VLMs as Kinematic Constraint Generators for Humanoid Loco-Manipulation

Alexandre Brown

Context & Problem

- Humanoid loco-manipulation typically relies on first learning a **General Motion Tracking (GMT) policy** that learns a motion prior.
- GMT policies mimic reference trajectories from joint states alone, **blind to the scene**. Kimodo generates motion from text but **lacks visual grounding.**

Research Question

- Can an off-the-shelf **VLM**, from a single ego-centric frame, **output Kinematic constraints** for Kimodo so a pretrained GMT policy acts **scene-aware**?

Approach

- We built a **10-task sim benchmark** (5 vision, 5 text). Qwen 2.5 VL predicts kinematic constraints, Kimodo generates the trajectory, and a pretrained GMT policy tracks it on **Unitree G1 in Isaac Lab**.

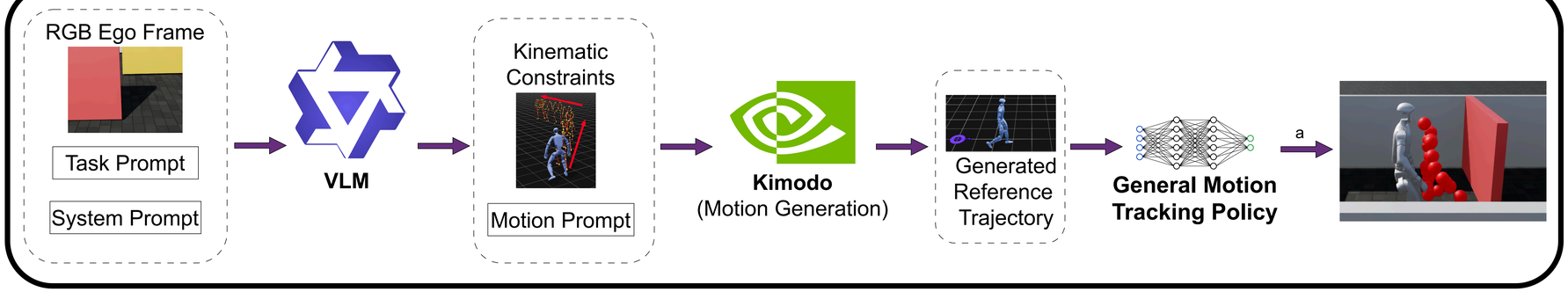
Results

- **Vision tasks:** +17% task progress on average.
- **Text tasks:** VLMs hurt performance.
- Different failure modes (bad VLM grounding, bad Kimodo motion, bad GMT tracking).

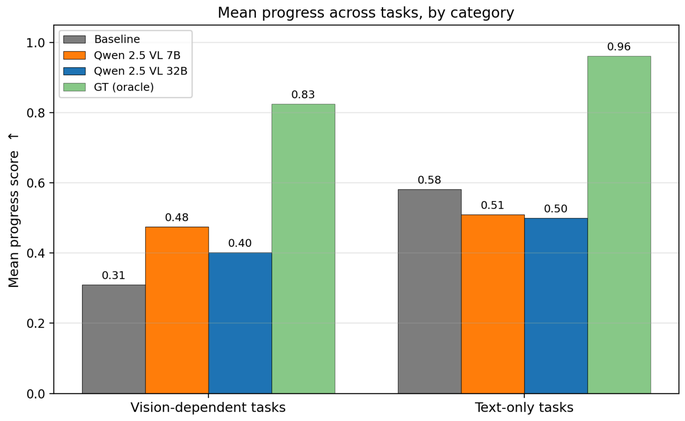
Future Work

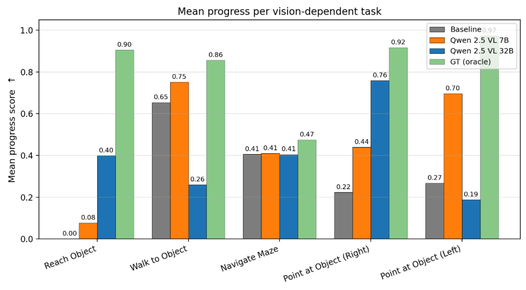
- Fine-tune or better in-context learning for the VLM.
- Natively multi-modal GMT policies to close the gap.

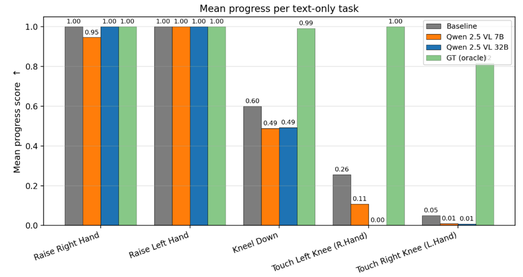
Method



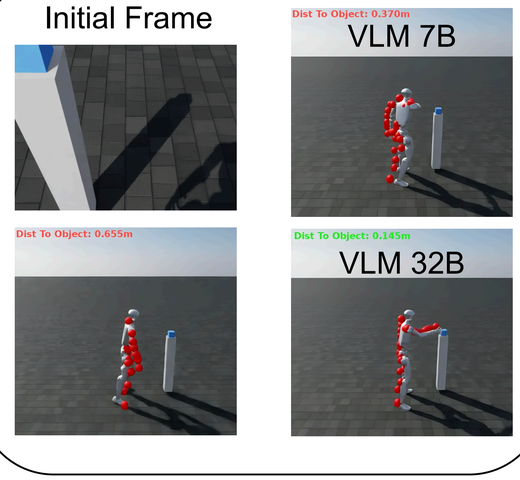
Results



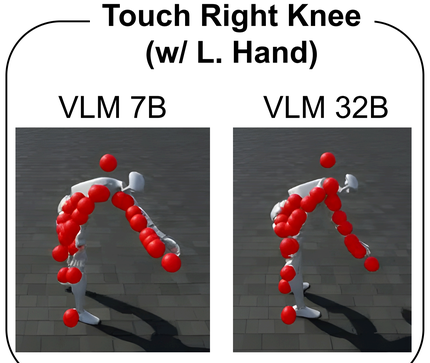




Point At Object With Right hand



Touch Right Knee (w/ L. Hand)



Failure Modes

