Constrained Multi-Objective Post-Training for Vision-Language-Action Models

Cody Hill[1], Zain Ulabedeen Farhat[2], Yue Wang[2], and Mubarak Shah[2]

[1]Mississippi State University, [2]University of Central Florida

National Science Foundation

Center for Research in Computer Vision

Mississippi State University

Institute of Artificial Intelligence

Introduction

Goal

We propose the Constrained Multi-Objective Group Relative Policy Optimization (CMO-GRPO) algorithm, which extends GRPO with preference conditioning and primal-dual constraint enforcement.

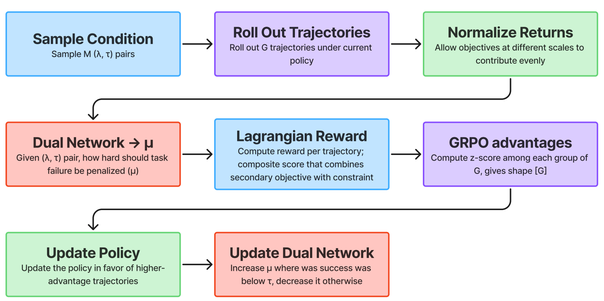
- Fine-tune pre-trained Vision-Language-Action (VLA)
- VLA model must Pareto optimize secondary objectives
- Secondary objectives subject to hard task-success floor

Motivation

- Success-only post training ignores deployment costs like energy, jerk, and collision risk.
- We want one policy π that respects a hard success floor and provides a tunable tradeoff dial at inference

Method

- Success floor τ enforced via Lagrangian duality
- Secondary objectives optimized beneath floor
- Preference conditioning: one policy serves a whole family of trade-offs, selected by λ.
- Group-relative advantages computed within (λ, τ) groups

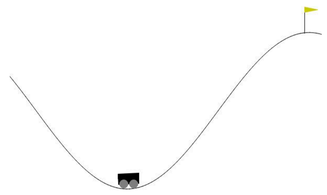


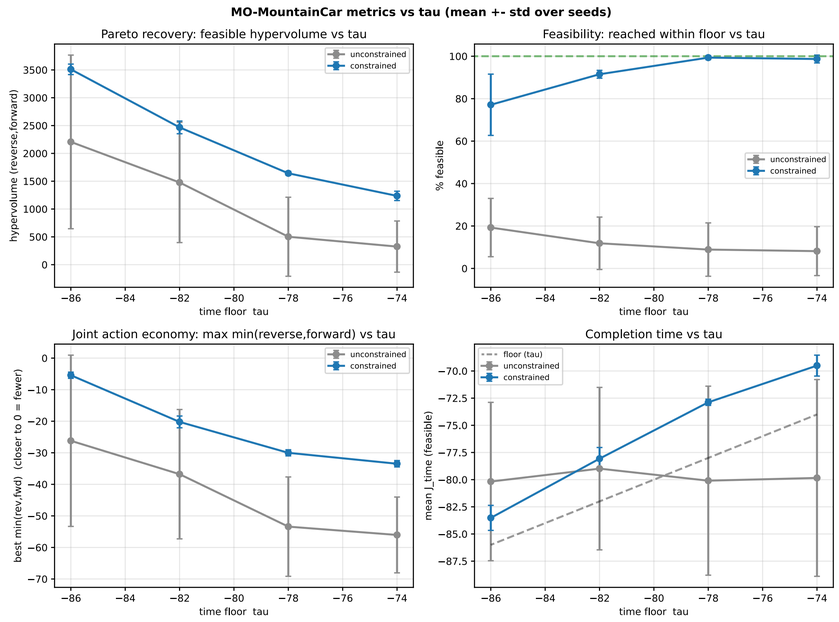
Results

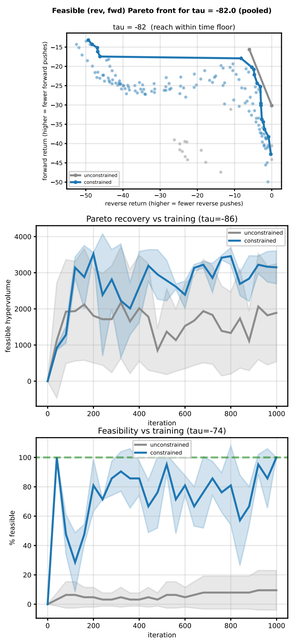
We chose to validate our algorithm on progressively difficult environments.

MO-MountainCar

- Constrained Objective: Time
- Secondary Objectives: Forward & Reverse
- Constrained Run: 77-99% floor satisfaction
- Unconstrained Run: 8-19% floor satisfaction



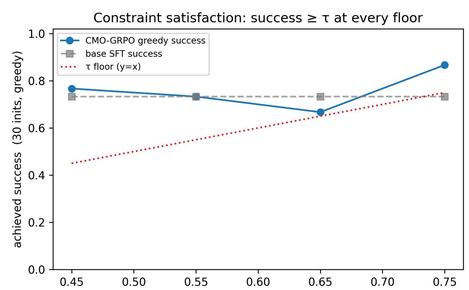




Application: LIBERO

Using MiniVLA, we performed supervised fine-tuning on the four LIBERO suites.

- τ-sweep holds the success floor at a 100% constraint satisfaction rate
- Balanced preference vertex matches base success and has lower energy



Fine-tuned Results

Suite	Avg. Success
Spatial	72%
Object	70%
Goal	75%
Long	40%

Key Finding

- Standard LIBERO only affords one steerable secondary objective.
- A PCA of achievable objective vectors gives ≈2.85 effective dimensions.
- Energy is the only axis with room to move, meaning the benchmark cannot stage a genuine 3+ way multi-objective competition.
- This is a property of the benchmark, not our method. It requires that we change how we evaluate multi-objective VLA post-training.

Conclusions + Future Work

- CMO-GRPO enforces hard success floors while steering secondary objectives.
- The limitations of standard LIBERO motivates the search for richer environments
- Next: LIBERO-Safety with graded collision/energy objectives, and competence-only SFT to obtain checkpoints the method can then steer.



Acknowledgements

Conducted in the University of Central Florida Computer Vision Research Experience for Undergraduates (Dr. Mubarak Shah, Dr. Niels da Vitoria Lobo, and Melody Halbert), advised by Dr. Yue Wang, with graduate mentor Zain Ulabedeen Farhat.