Implicit Safety Alignment from Crowd Preferences

Qian Lin, Daniel S. Brown
Kahlert School of Computing, University of Utah

ICML 2026

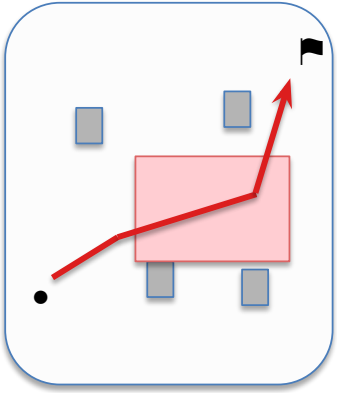
Motivation / Problem Setting

From crowd preferences, we extract shared safety and transfer to new tasks

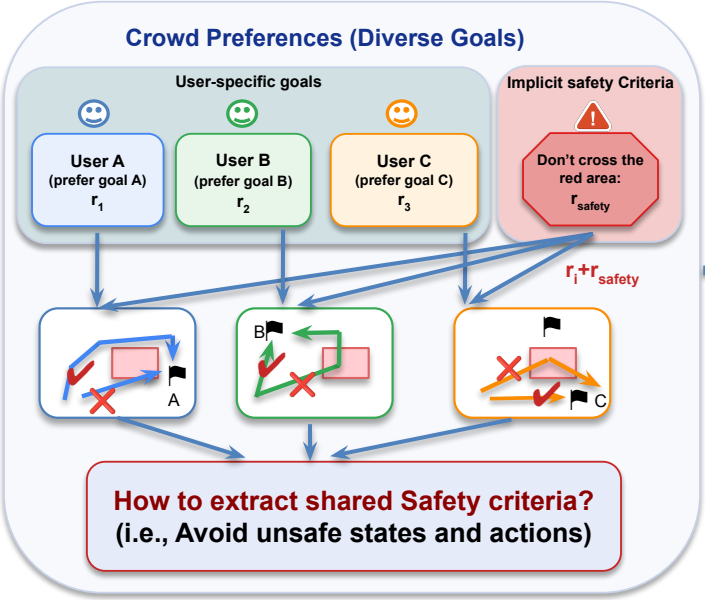
Problem:
Using safety-agnostic task reward leads to unsafe behavior

Solution: Crowd preferences from diverse users with shared safety

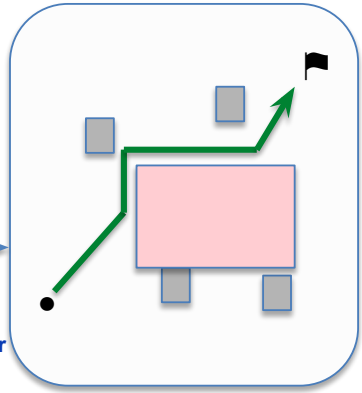
Goal: Combining shared safety with task reward yields safe behavior



Policy optimized by task reward only
→ High task reward, but unsafe



Transfer Safety criteria



Policy with task reward + shared safety criteria
→ High task reward and safe

● Start ⚑ Goal ■ Obstacle ▫ Danger Zone --- Unsafe Trajectory --- Safe Trajectory

Why Vanilla RLHF Fails

Natural solution: reward combination

Crowd Preferences

↓

Learn Reward Model r_{pref}

↓

Optimize

$r = (1-\omega)\, r_{task} + \omega\, r_{pref}$

- r_{task} :Safety-agnostic
- r_{pref} :Safety-aware

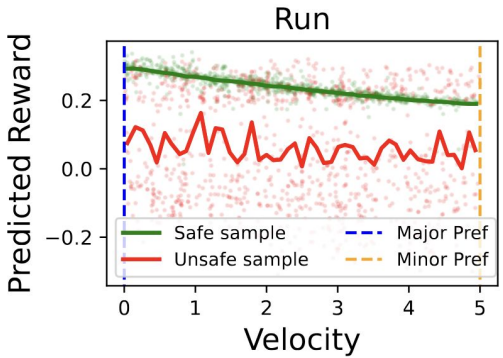
↓

Entangles shared safety r_{safety} with user-specific component r_i

Two major limitations

1. Preference imbalance bias

- Safe trajectories get higher reward
- Majority preferences get over-rewarded

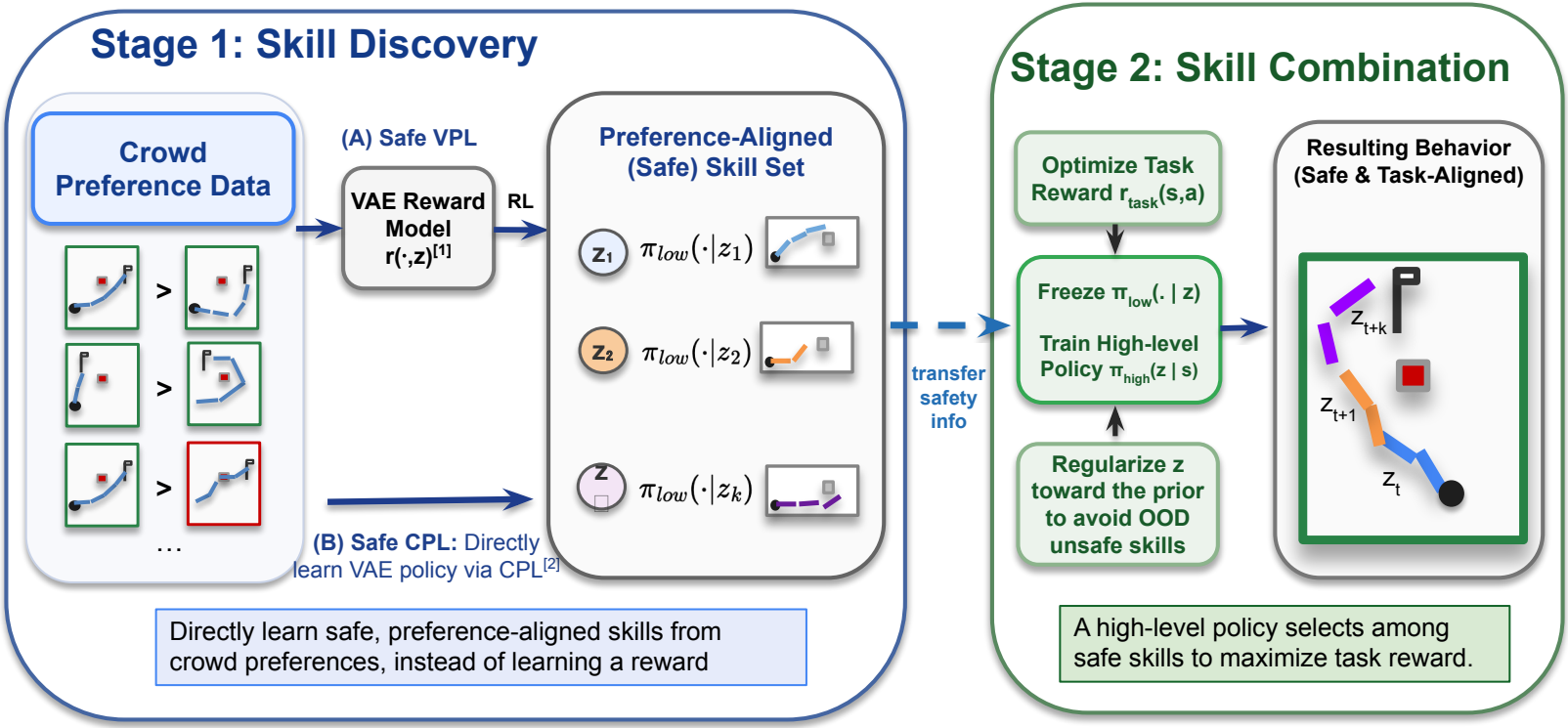


2. Sensitive to ω

- Small ω → unsafe
- Large ω → poor task performance

Our Method: Composing Safe Skills

Instead of combining rewards, we compose safe behaviors (skills)



key idea: the high-level policy solves the task, and the low-level skills provide safety.

1. Poddar, Sriyash, et al. "Personalizing reinforcement learning from human feedback with variational preference learning." NeurIPS 2024.
2. Hejna, Joey, et al. "Contrastive preference learning: Learning from human feedback without reinforcement learning." ICLR 2024.

Experimental Results

Safe skill composition lowers safety cost while keeping task performance comparable to oracle.

Evaluation Setup

Six safe RL environments
Reach, Run, Circle, Ant-vel, Swimmer-vel, HalfCheetah-vel

Proof-of-concept LLM-style evaluation
Avoid harmful responses

Preference data
Different goals, shared safety
(*e.g., users prefer different navigation targets but consistently avoid unsafe areas*)

Downstream Task
A safety-agnostic reward specifies an unseen goal

Metrics
- Normalized reward ↑
- Normalized safety cost ↓

Baselines
Task-only, SOPL, Oracle, Reward Combination (RC),

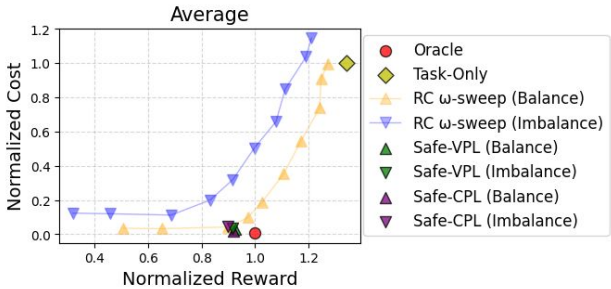
Ours: Safe-VPL / Safe-CPL

Main Results (Average over 6 Environments)

Stats	Oracle	Task-Only	SOPL	RC($\omega = 0.5$)	Safe-VPL	Safe-CPL
Norm Rew	1.00	1.46	1.04	0.82	0.93	0.92
Norm Cost	0.01	1.00	0.01	0.05	0.03	0.02

✖ Task-Only → severe safety violations
✖ Oracle / SOPL → strong, but require explicit safety reward or preference
✖ RC → struggles to balance safety and task performance
✓ Ours → comparable reward to Oracle / SOPL with ~97% lower safety cost than Task-Only

Sensitivity to preference imbalance



✖ RC is sensitive to the trade-off weight ω and degrades under preference imbalance.

✓ Under both balanced and imbalanced preferences, our methods stay on the low-cost frontier and are robust to imbalance.