Diversity Is Not All You Need: Training A Robust Cooperative Agent Needs Specialist Partners

Rujikorn Charakorn, Poramate Manoonpong, Nat Dilokthanakul

VISTEC SDU IT

Cross-play minimization (XP-min) generates diverse agents

$$\max_{\pi_A} J_{\texttt{XP-min}}(\pi_A, \mathcal{P}) = \overbrace{J_{\mathrm{SP}}(\pi_A)}^{\text{High SP return}} \overbrace{- \lambda_{\mathrm{XP}} J_{\mathrm{XP}}(\pi_A, \pi_+)}^{\text{Low XP return}}\ ;\ \forall \pi_A \in \mathcal{P},$$

$$\pi_+ = \underset{\pi_+ \in (\mathcal{P} \setminus \{\pi_A\})}{\operatorname{argmax}} J_{\mathrm{XP}}(\pi_A, \pi_+),$$

Mix-play regularization (MP-reg) reduces overfitness of XP-min agents but agents might lose specialization

$$\max_{\pi_A} J_{\texttt{MP-reg}}(\pi_A, \mathcal{P}) = \overbrace{J_{\mathrm{SP}}(\pi_A)}^{\text{High SP return}} \overbrace{- \lambda_{\mathrm{XP}} J_{\mathrm{XP}}(\pi_A, \pi_+)}^{\text{Low XP return}} \overbrace{+ J_{\mathrm{MP}}(\pi_A, \pi_+)}^{\text{High MP return}}\ ;\ \forall \pi_A \in \mathcal{P},$$

Core result #1: Three measures representing a population's quality

$x = f(\tau)$ is a characteristic of a trajectory τ given a characteristic function f

Diversity $$\mathcal{D}(\mathcal{P}) := H(X) = -\sum_x P(x) \log P(x) = -\sum_x \mathbb{E}_{\pi}[P(x|\pi)] \log \left(\mathbb{E}_{\pi}[P(x|\pi)]\right),$$

Entropy of the trajectory characteristic of an entire population

Specialization $$\mathcal{S}(\mathcal{P}) := -\mathbb{E}_{\pi}\left[H(X \mid \Pi = \pi)\right] = -H(X \mid \Pi),$$

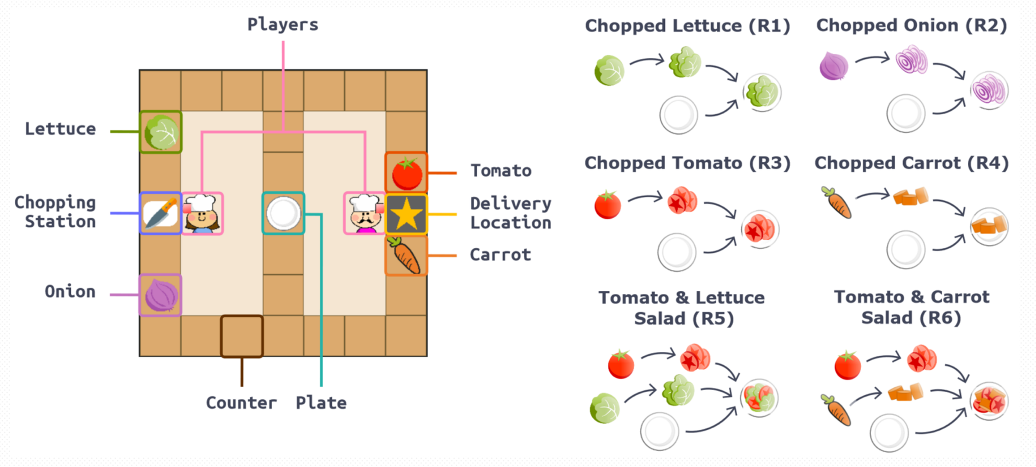
$$H(X \mid \Pi = \pi) = -\sum_x P(x|\pi) \log P(x|\pi)$$

Negative expected entropy of the characteristic of each joint policy

Overfitness $$\mathcal{O}(\mathcal{P}) = 1 - R(\pi_G^*; \mathcal{P})$$

The compliment of expected success rate (R) of the joint policies in the population when matched with an oracle generalist (OG)

Evaluation environment: Multi-recipe Overcooked



$x = f(\tau)$ is a one-hot vector representing the completed recipe of a trajectory τ

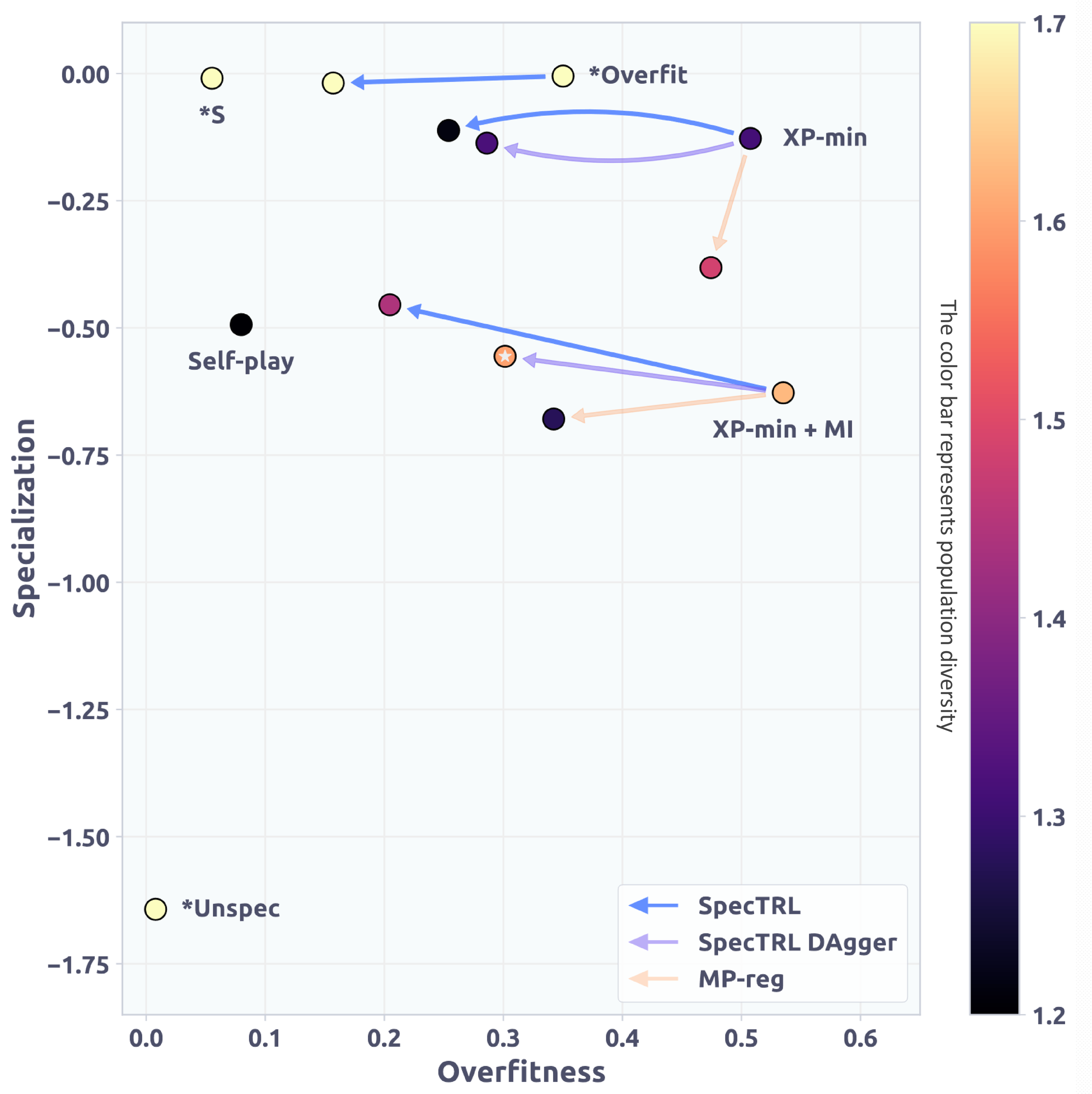
Core result #2: Unspecialized or overfit partners induce less robust generalist agents

Populations	$\mathcal{D}(\mathcal{P})$	$\mathcal{S}(\mathcal{P})$	$\mathcal{O}(\mathcal{P})$	$\mathcal{R}(\pi_G, \mathcal{P}_{\text{test}})$
(•) $\mathcal{P}^*_S$	1.79	−0.01	0.06	**0.81** ± 0.05
(•) $\mathcal{P}^*_{\text{unspec}}$	1.72	**−1.64**	0.01	0.49 ± 0.07
(•) $\mathcal{P}^*_{\text{overfit}}$	1.79	−0.01	**0.35**	0.73 ± 0.01
(•) $\mathcal{P}_{\texttt{XP-min}}$	1.31	−0.12	0.51	0.49 ± 0.02

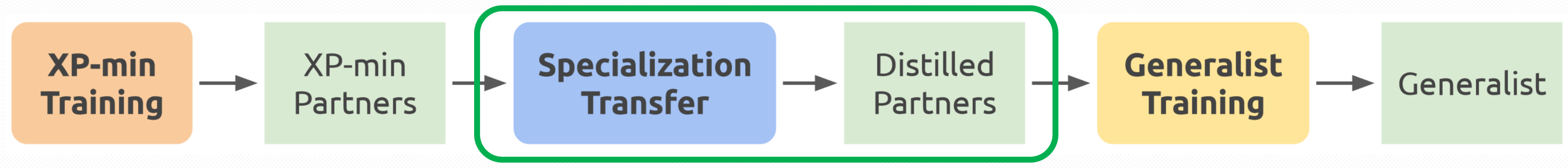
We show that partners' specialization, in addition to diversity, is crucial for training a robust cooperative agent.

SpecTRL reduces overfitness of XP-min partners that already have good diversity and specialization

The partner quality landscape



Proposed method: Specialization Transfer



SpecTRL: Specialization transfer with reinforcement learning

$$J_{\texttt{SpecTRL}}(\pi_{A'}) = \sum_{i=1}^{i=N} J(\pi^i_{A'}, \pi^{-i}_{A})$$

distilling via the reward maximization objective incentivizes the distilled partners to "nudge" the source partners to perform cooperative behaviors

SpecTRL DAgger: Adding DAgger to SpecTRL

$$J_{\texttt{SpecTRL DAgger}}(\pi_{A'}) = \sum_{i=1}^{i=N} J(\pi^i_{A'}, \pi^{-i}_{A}) + \lambda_{\mathrm{DAgger}} \mathcal{L}_{\mathrm{DAgger}}(\pi^i_{A'}),$$

$$\mathcal{L}_{\mathrm{DAgger}}(\pi^i_{A'}) = -\mathbb{E}_{\tau^i_t \sim \rho(\pi^i_{A'}, \pi^{-i}_{A})} \log \pi^i_{A'}(\hat{a}^i_t | \tau^i_t),$$

Useful for stabilizing the distillation process by directly transferring the knowledge from the source policy

Core result #3: SpecTRL removes overfitness

Populations	$\mathcal{D}(\mathcal{P})\uparrow$	$\mathcal{S}(\mathcal{P})\uparrow$	$\mathcal{O}(\mathcal{P})\downarrow$	$\mathcal{R}(\pi_G, \mathcal{P}_{\text{test}})\uparrow$
*S	1.79	−0.01	0.06	0.82 ± 0.05
*Overfit	1.78	0.00	0.35	0.74 ± 0.02
*Unspec	1.72	−1.64	0.01	0.49 ± 0.08
[*Overfit] + SpecTRL	1.79 (≈)	−0.01 (≈)	0.16 (⇓ 0.19)	0.78 ± 0.01 (⇑ 0.04)

Core result #4: SpecTRL DAgger removes overfitness while maintaining diversity when applied to XP-min population

Populations	$\mathcal{D}(\mathcal{P})\uparrow$	$\mathcal{S}(\mathcal{P})\uparrow$	$\mathcal{O}(\mathcal{P})\downarrow$	$\mathcal{R}(\pi_G, \mathcal{P}_{\text{test}})\uparrow$
[XP-min] + MI	1.63 ± 0.03 (⇑ 0.32)	−0.63 ± 0.12 (⇓ 0.50)	0.54 ± 0.06 (≈)	0.61 ± 0.02 (≈)
[XP-min + MI] + SpecTRL	1.44 ± 0.04 (⇓ 0.19)	−0.45 ± 0.06 (⇑ 0.18)	0.20 ± 0.06 (⇓ 0.34)	0.62 ± 0.01 (≈)
[XP-min + MI] + SpecTRL DAgger	1.60 ± 0.02 (≈)	−0.56 ± 0.08 (≈)	0.30 ± 0.03 (⇓ 0.24)	0.64 ± 0.01 (⇑ 0.03)