Learning to Cooperate with Humans using Generative Agents

Yancheng Liang, Daphne Chen, Abhishek Gupta, Simon S. Du*, Natasha Jaques*
University of Washington
{yancheng, daphc, abhgupta, ssdu, nj}@cs.washington.edu

Abstract

Training agents that can coordinate zero-shot with humans is a key mission in multi-agent reinforcement learning (MARL). Current algorithms focus on training simulated human partner policies which are then used to train a Cooperator agent. The simulated human is produced either through behavior cloning over a human dataset, or by using MARL to create a population of simulated agents. However, these approaches often struggle to produce an effective Cooperator since the simulated humans fail to cover the diverse strategies employed by people in the real world. We show that *learning a generative model of human partners* can effectively address this issue. Our model learns a latent variable representation of the human that can be regarded as encoding the human's unique strategy, intention, experience, or style. This generative model can be flexibly trained from any (human or neural policy) agent interaction data, unifying approaches proposed in prior work. By sampling from the latent space, we can use the generative model to produce different partners to train Cooperator agents. We evaluate our method—**G**enerative **A**gent **M**odeling for **M**ulti-agent **A**daptation (**GAMMA**)—on Overcooked, a challenging cooperative cooking game that has become a standard benchmark for zero-shot coordination. We conduct an evaluation with real human teammates, and the results show that **GAMMA** consistently improves performance, whether the generative model is trained on simulated populations or human datasets. Further, we propose a method for posterior sampling from the generative model that is biased towards the human data, enabling us to efficiently improve performance with only a small amount of expensive human interaction data. [1]

1 Introduction

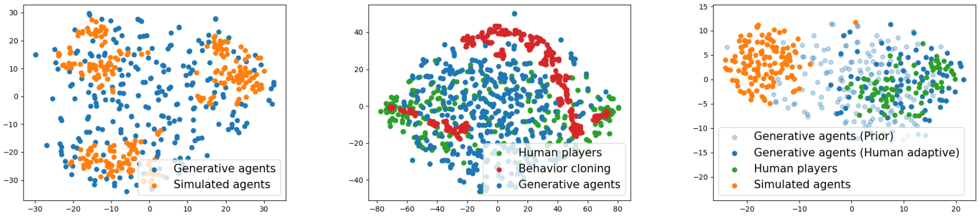
While being cooperative is a notable characteristic of human intelligence [7, 25], training an artificial agent that cooperates well with humans poses a significant challenge. Human behaviors are uncertain and diverse, encompassing a wide range of preferences, abilities, and intentions. While humans can rapidly adapt to different partners, AI agents are particularly poor at generalizing to working with a novel human partner [1, 22].

There are two approaches to tackling this problem. The first is to use data of human-human interactions [1]. However, in many situations the amount of human-human data available is far less than the amount of data required by data-hungry sequential decision making algorithms. This suggests using simulated training partners as a different approach. However traditional MARL approaches

[1] See our website for human-AI study videos and an interactive demo. The training code is also available.

[2] Points are the latent vectors. Taking the population of simulated agents (e.g., the FCP population with 8 FCP agents) as an example, a point is generated by: 1) sampling an agent in the FCP population; 2) using the agent to generate an episode with self-play; 3) using the VAE encoder to encode the episode into a latent vector and mapping the latent vector to the 2D space using t-SNE.

NeurIPS 2024 Workshop on Behavioral Machine Learning.



(a) Simulated data.

(b) Human data.

(c) Adaptive sampling.

Figure 1: The latent space covered by different methods.[2]For either simulated data or human data, the generative agents produced by **GAMMA** cover a larger strategy space. Generative models can provide novel agents by interpolating the agents in the simulated population (a). On human data (b), the human proxy model only covers a subset of all human behavior patterns, while the generative model can capture the diversity in the data. We can also control the latent space sampling (c) to model a target population of agents (e.g., human coordinators).

like self-play may fail to coordinate effectively with novel humans [1, 22] because the self-play agent merely follows a convention with a copy of itself [6] and fails to adapt to a human's novel strategy.

To tackle these challenges, a popular approach to solving the distribution shift problem in human-AI cooperation has become training a Cooperator agent against a *population* of simulated agents rather than just itself [2, 12, 14, 18, 22, 28, 30]. However, the bottom line remains that truly covering the space of strategies with discrete samples of strategies can quickly become untenable. As we move towards more complex real-world tasks, representing every strategy by an individual agent in the population becomes computationally difficult. The above two approaches point to the dual problems of 1) human-only data being expensive, and 2) synthetic-only data lacking coverage over human behaviors.

We propose our full method **GAMMA**: **G**enerative **A**gent **M**odeling for **M**ulti-agent **A**daptation. The idea is to learn a generative model to generate a diverse range of partners by interpolating on the interaction data it has been trained on, as shown in Figure 1. We propose to leverage variational autoencoder (VAE) [11] techniques to train the model. The latent variable z is designed to capture human diversity by encoding information about that partner's unique style or skill levels. By sampling z, the decoder can be used to generate the different partners, which are used to train the Cooperator. With the availability of a controllable latent encoder, we further propose a human-adaptive sampling method that enables training Cooperators that are more targeted to cooperate with real human coordinators. Thus we can incorporate a small amount of human-provided data in a economical way.

We test our method using Overcooked [1], a collaborative multiplayer cooking game requiring close coordination between two partners to successfully prepare recipes. We conducted our evaluation through real human experiments. By having the train agents play in real-time against real human players on a simplified simulated Overcooked platform, we find that **GAMMA** improves the performance of state-of-the-art coordination methods which rely on simulated populations of agents [18, 22], or on human data [1]. We also find that humans can adapt very quickly and learn through playing (see Appendix **??**).

2 Related Work

Zero-shot coordination. Building cooperative agents for novel partners is a long-standing problem of AI [21], known as zero-shot coordination [10]. One approach to this problem relies on collecting and learning from human-human data [1], but this is challenging due to limited scale, and the heterogeneity, uncertainty and suboptimality [8, 16] of humans. Instead, another line of work focuses on training a simulated population of partner policies which can be used to train a single robust Cooperator agent (e.g. [22]). As such, many papers have focused on strategies for making this population more diverse [2, 3, 14, 17, 18, 23, 24, 26, 28, 30]. In contrast, we propose a novel approach to *modeling* this population with a generative model.

Inferring the partner type. Our work is also related to the many works (e.g. [5, 9, 15, 20, 27]), which train a latent variable model to learn agent representations from interaction data of different agents, and then infer a cooperation partner's latent vector and use it to condition a multi-agent RL policy. While our work also infers hidden context about partner type, our aims are different; we use a generative model to simulate novel partners during training time as a way to train a Cooperator to be robust enough for zero-shot coordination with real humans.

3 GAMMA: Generative Agent Modeling for Multi-agent Adaptation

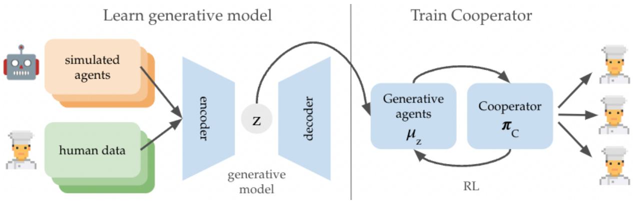


Figure 2: Overview of the method for **GAMMA**. The generative model learns a latent distribution from either simulated or human data. Sampling partners from the generative model enables training a robust Cooperator that can coordinate with a variety of different humans.

Learning Generative Models of Partner Behavior. We assume access to a dataset of trajectories $\mathcal{D}_{\text{coordination}} = \{\tau_i\}_{i=1}^N$, where each trajectory $\tau = \{(s_t, a_t, b_t)\}_{t=0}^T$ (here s_t, a_t, b_t represent the state, the agent's own action, and the other agent's action, respectively) is a sequence of multi-agent coordination behaviors, derived from either human playing records or paired simulated agents. We develop a seq-to-seq variant of the Variational Autoencoder (VAE) [5, 11] to model the dataset $\mathcal{D}$. An approximate posterior (encoder) $q(z \mid \tau; \phi)$ identifies the agent style from the trajectory. The decoder $p(a_t \mid z, \tau_{0..t-1}; \theta)$ uses the agent's own past experience and the latent variable z to predict the agent's next action. The generative model is trained using an evidence lower bound (ELBO) loss **??**.

Training a Cooperator with Generative Coordination Models. The trained generative agent model $p(a_t \mid z, \tau_{0..t-1}; \theta)$ can now be used as a generator of partner policies μ_z to train our Cooperator agent. Notably, the generative model can sample a landscape of agents going beyond the quantity and diversity of the training data. Then we use PPO [19] to optimize π_C against these training partners, following the objective $J(\pi_C) = \mathbb{E}_{z \sim p(z)} \left[V(\pi, \mu_z) \right]$. Importantly, the Cooperator π_C is not trained with imitation learning, but is rather trained with reinforcement learning. The generative model from Section 3 is simply used to provide the quantity and diversity of agents.

Targeted GAMMA using Human-Adaptive Sampling and Fine-tuning. The above approach does not make use of human specific data if available. The key insight is that the latent space afforded by our generative model provides the ability to do *controllable sampling* from any latent distribution. In particular, when coordinating with real human partners, a human-centered latent distribution $p_h(z)$ can be estimated by $p_h(z) = \mathcal{N}(\bar{z}, I)$ where $\bar{z} = \mathbb{E}_{\tau_h \in D_h} \left[\mathbb{E}_{z \in q(\cdot|\tau_h)}[z] \right]$. Then a targeted Cooperator can be trained by maximizing $J(\pi_C) = \mathbb{E}_{z \sim p_h(z)} \left[V(\pi, \mu_z) \right]$.

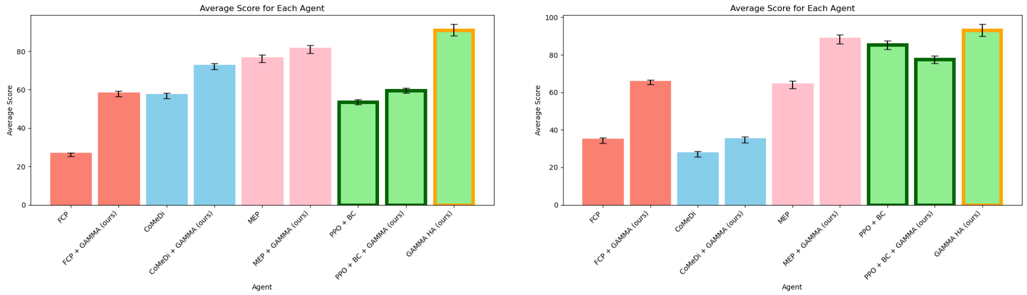
Human-adaptive sampling makes the model focus less on adaptation to irrelevant synthetic partners and more on "human-centric" partners. As our experiments will show, this approach **Human-Adapative (HA)** outperforms training a partner simulator using only human data. To better capture human data, we also perform some fine-tuning on the encoder and decoder using human data.

4 Experiments

We use the Overcooked environment [1] as a popular benchmark for prior work on human-AI cooperation [1, 18, 22, 27, 28]. A detailed description of Overcooked can be found in Appendix **??**.

Hypothesis: Generalizing to novel humans. Will **GAMMA** generalize to playing with novel human partners more effectively than prior work, as measured by the score in the cooperative game? What will humans' self-reported ratings reveal about their preferences over different methods?

Baselines and Data We include two state-of-the-art approaches for creating a population of simulated agents, **Fictitious Co-Play (FCP)** [22] and **Cross-play Optimized, Mixed-play Enforced Diversity (CoMeDi)** [18]. We then train the generative models **FCP+GAMMA** and **CoMeDi+GAMMA** on the simulated agent population of both FCP and CoMeDi. We also use PPO-BC [1] as the baseline which incorporates human data, by training a BC partner and an RL Cooperator. In this work, we



(a) Counter circuit layout

(b) Multi-strategy Counter

Figure 3: Performance of different agents when played with real humans. Error bars [4] use the Standard Error of the Mean (SE) for statistical significance ($p < 0.05$). Methods trained on human data are shown in green. Whether training with simulated or human data, **GAMMA** shows consistent, statistically significant advantages over the baselines. **GAMMA-HA** is able to efficiently use the real human dataset to learn a better sampling of its latent space, achieving the best performance when cooperating with real humans.

assume the human dataset contains 20 to 50 trajectories. We will test whether replacing the BC agent with our generative model (**PPO-BC-GAMMA**) trained on the human data provides better performance. Since we assume the amount of human data is limited, we also fine-tune a generative model pretrained from the simulated agent population with human data, and apply **Human-Adaptive (HA)** to train the Cooperator (**GAMMA-HA**).

Human Evaluation. We run a user study with real human players in order to determine which method can most effectively coordinate with humans, and which method is preferred in human ratings (Hypothesis). We conducted a study with 80 users recruited via online crowdsourcing from Prolific. Our study follows guidelines set by an IRB protocol. During the study, each user is instructed to play multiple rounds of Overcooked with a partner via a web interface, where in each round the partner is an agent following one of the 8 policies, in randomized order. We trained 5 random seeds per agent, and used a different randomly-selected seed for each of the 8 game rounds. Each game lasts for 60 seconds. After each round, the user answers Likert scale survey questions [13] to rate their experience playing with the agent. At the end of the 8 rounds, humans also answer qualitative questions about the performance of the agents. We conduct a qualitative analysis to understand which factors most heavily influence overall performance and users' preferences when playing with real humans.

5 Results

As described in Section 4, we evaluate all agents described before in Section 4 against the novel human players, and plot the cooperative scores achieved by each agent-human team in Figure 3.

We find that **GAMMA** offers consistent, significant performance improvements over prior techniques for training against simulated populations (**FCP, CoMeDi**). **GAMMA** does not perform strongly when training on the CoMeDi population in the second layout, but this is because the CoMeDi method failed to generate a robust population that covered the diversity of strategies in the second layout. Comparing methods that make use of human data reveals some interesting findings. Modeling the human data with the generative model (**PPO+BC+GAMMA**) only provides performance improvements half the time. However, combining simulated and human data with **Human Adaptive GAMMA** provides significantly higher performance in both layouts, surpassing state-of-the-art zero-shot coordination techniques.

In the human study, we find that **adaptation to the human partner** was a core theme distinguishing agents that humans liked. Participants reported that the **FCP + GAMMA** agent demonstrated an ability to learn from the user's actions, such as mimicking the user's strategy of placing onions on the table to save time. This adaptive behavior was positively received by a study participant: "I noticed that once I started to put back onions on the table that it did the same as I wanted to save time rather than going back for onions 3 times for soup. I thought it was interesting that it learned about my behavior." We provide more subjective ratings from the human players as well in Appendix G.2.

6 Conclusion

In this work, we propose **GAMMA**, a novel approach to training a coordinator agent by using generative models to produce training partner agents. We conduct a comprehensive analysis using data from a study with real human cooperation partners, and show **GAMMA** outperforms baselines over both subjective human ratings and quantitative measurements of cooperation performance. We also provide a new perspective to compare different populations under the latent space of a generative model, showing how the simulated populations may not provide sufficient coverage of the range of human players.

Limitations. As shown by the performance of GAMMA+CoMeDi on *Multi-Strategy Counter*, obtaining good performance with our approach depends on having a reasonably diverse amount of cooperation data to train the model. If the quality of the simulated population data is too low, the approach can fail to provide significant benefits.

In this work, our human studies recruit participants from Prolific, which may not be representative of broader populations. Additionally, our human dataset is limited, which could reduce the diversity of strategies and force participants to adapt to strategies that the Cooperators are already familiar with.

We focus on the two-player setting in this study following prior work [22, 28, 30] because it is a first step toward enabling an AI assistant that could help a human with a particular task. Scaling up to more agents would exponentially increase the dataset size with our current techniques. Therefore, better sampling techniques are needed to address this issue.

Future work. Several potential directions are interesting for future work: 1) In this work, the amount of human data is limited, which restricts the performance of the generative model that learns human data from scratch. 2) An orthogonal direction is to condition the policy of the Cooperator on the embedding of the partner policy. We provide some preliminary results in Figure 12.

Social impact. Our work focuses on how to train AI agents that can effectively cooperate with diverse humans to assist them with tasks. We believe this is a critical component of eventually enabling assistive robots that could operate in human environments to assist the elderly or disabled to live more comfortably, or reduce the burden of domestic labour for all people.

Acknowledgment

SSD acknowledges the support of NSF IIS 2110170, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, NSF CCF 2019844, and NSF IIS 2229881.

References

[1] M. Carroll, R. Shah, M. K. Ho, T. L. Griffiths, S. A. Seshia, P. Abbeel, and A. D. Dragan. On the utility of learning about humans for human-ai coordination. *CoRR*, abs/1910.05789, 2019. URL `http://arxiv.org/abs/1910.05789`.

[2] R. Charakorn, P. Manoonpong, and N. Dilokthanakul. Generating diverse cooperative agents by learning incompatible policies. In *The Eleventh International Conference on Learning Representations*, 2022.

[3] B. Cui, A. Lupu, S. Sokota, H. Hu, D. J. Wu, and J. N. Foerster. Adversarial diversity in hanabi. In *The Eleventh International Conference on Learning Representations*, 2022.

[4] G. Cumming, F. Fidler, and D. L. Vaux. Error bars in experimental biology. *The Journal of cell biology*, 177(1):7–11, 2007.

[5] A. Grover, M. Al-Shedivat, J. Gupta, Y. Burda, and H. Edwards. Learning policy representations in multiagent systems. In *International conference on machine learning*, pages 1802–1811. PMLR, 2018.

[6] N. Grupen, N. Jaques, B. Kim, and S. Omidshafiei. Concept-based understanding of emergent multi-agent behavior. In *Deep Reinforcement Learning Workshop NeurIPS 2022*, 2022.

[7] J. Henrich. The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter. In *The secret of our success*. princeton University press, 2015.

[8] S. J. Hoch and G. F. Loewenstein. Time-inconsistent preferences and consumer self-control. *Journal of consumer research*, 17(4):492–507, 1991.

[9] J. Hong, A. Dragan, and S. Levine. Learning to influence human behavior with offline reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[10] H. Hu, A. Lerer, A. Peysakhovich, and J. Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.

[11] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In Y. Bengio and Y. LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6114`.

[12] Y. Li, S. Zhang, J. Sun, Y. Du, Y. Wen, X. Wang, and W. Pan. Cooperative open-ended learning framework for zero-shot coordination. In *International Conference on Machine Learning*, pages 20470–20484. PMLR, 2023.

[13] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.

[14] A. Lupu, B. Cui, H. Hu, and J. Foerster. Trajectory diversity for zero-shot coordination. In *International conference on machine learning*, pages 7204–7213. PMLR, 2021.

[15] G. Papoudakis, F. Christianos, and S. Albrecht. Agent modelling under partial observability for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 19210–19222, 2021.

[16] J. W. Pratt. Risk aversion in the small and in the large. In *Uncertainty in economics*, pages 59–79. Elsevier, 1978.

[17] J. K. Pugh, L. B. Soros, and K. O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

[18] B. Sarkar, A. Shih, and D. Sadigh. Diverse conventions for human-ai collaboration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[20] M. Shum, M. Kleiman-Weiner, M. L. Littman, and J. B. Tenenbaum. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6163–6170, 2019.

[21] P. Stone, G. Kaminka, S. Kraus, and J. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1504–1509, 2010.

[22] D. Strouse, K. McKee, M. Botvinick, E. Hughes, and R. Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515, 2021.

[23] Z. Tang, C. Yu, B. Chen, H. Xu, X. Wang, F. Fang, S. S. Du, Y. Wang, and Y. Wu. Discovering diverse multi-agent strategic behavior via reward randomization. In *International Conference on Learning Representations*, 2020.

[24] B. Tjanaka, M. C. Fontaine, J. Togelius, and S. Nikolaidis. Approximating gradients for differentiable quality diversity in reinforcement learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1102–1111, 2022.

[25] M. Tomasello. *Why we cooperate*. MIT press, 2009.

[26] S. Wu, J. Yao, H. Fu, Y. Tian, C. Qian, Y. Yang, Q. Fu, and Y. Wei. Quality-similar diversity via population based reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[27] S. A. Wu, R. E. Wang, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, and M. Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021.

[28] C. Yu, J. Gao, W. Liu, B. Xu, H. Tang, J. Yang, Y. Wang, and Y. Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *The Eleventh International Conference on Learning Representations*, 2022.

[29] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

[30] R. Zhao, J. Song, Y. Yuan, H. Hu, Y. Gao, Y. Wu, Z. Sun, and W. Yang. Maximum entropy population-based training for zero-shot human-ai coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6145–6153, 2023.

A Reproducibility

Our demo website is https://sites.google.com/view/human-ai-gamma-2024/ and contains the code and more experiment results. We also provide information about the implementation details B and hyperparameters used in our experiments D to help reproduce our results.

B Implementation details

Generative models. The dataset used to train the VAE model contains the joint trajectories of two players. For the simulated agent population $\{\pi_1, ..., \pi_N\}$, we create this dataset by evenly sampling $\pi_i \times \pi_j$ to generate the trajectories. A simulated dataset contains 100K joint trajectories. To train a VAE on it, the dataset is split into a training dataset with 70% data and a validation dataset with the rest of 30% data. To compute the ELBO loss **??**, the trajectories are truncated to length 100 for better optimization for the recurrent module. A linear scheduling of the KL penalty coefficient β is adopted to control a target value for the KL divergence of the posterior distribution. The target KL value is set to 7 for the layout *Forced Coordination* over the CoMeDi population. All other VAE models are chosen by a target KL value of 32.

Train Cooperator agents. On Overcooked, the Cooperator is trained by PPO [19]. We based our implementations on HSP [28]. Reward shaping for dish and soup pick-up is used for the first 100M steps to encourage exploration. All results are reported with averaged episode reward and the standard error of at least 5 seeds.

Simulated agent populations. We use MAPPO [29] to create the FCP [22] agent populations. Eight self-play agents are trained and three checkpoints for each agent are added to the population, making the population size 8×3. For the CoMeDi agent population [18], we download the population proposed by the authors [3] for the original five layouts [1]. The population size is 8 for CoMeDi. For our custom layout *Multi-strategy Counter*, we use CoMeDi's official implementation and keep the population size the same.

C Human Dataset

For the human dataset in the original Overcooked paper [1], their open-sourced dataset contains "16 joint human-human trajectories for Cramped Room environment, 17 for Asymmetric Advantages, 16 for Coordination Ring, 12 for Forced Coordination, and 15 for Counter Circuit.." with length of $T \approx 1200$. In Multi-strategy Counter, we collect 38 trajectories with length $T \approx 400$ which is closer to the actual episode length during training.

D Hyperparameters

We use MAPPO to train our Cooperator agent. The architectures and hyperparamers are fixed throughout all layouts. All policy networks follow the same structure where an RNN (we use GRU) is followed by a CNN.

The generative model follows a similar architecture to the policy model. An encoder head and a decoder head are used to produce variational posterior and action reconstruction predictions from the representations.

E Computational Resources

We conducted our main experiments on clusters of AMD EPYC 64-Core Processor and NVIDIA A40/L40. It takes about one day to train one Cooperator agent. The main experiments takes about 3600 GPU hours. We do some preliminary experiments to search for the best hyperparameters and training frameworks.

[3] https://github.com/Stanford-ILIAD/Diverse-Conventions/tree/master

hyperparameter	value
CNN kernels	[3, 3], [3, 3], [3, 3]
CNN channels	[32, 64, 32]
hidden layer size	[64]
recurrent layer size	64
activation function	ReLU
weight decay	0
environment steps	100M (simulated data) or 150M (human data)
parallel environments	200
episode length	400
PPO batch size	$2 \times 200 \times 400$
PPO epoch	15
PPO learning rate	0.0005
Generalized Advantage Estimator (GAE) λ	0.95
discounting factor γ	0.99

Table 1: Hyperparameters for policy models

hyperparameter	value
CNN kernels	[3, 3], [3, 3], [3, 3]
CNN channels	[32, 64, 32]
hidden layer size	[256]
recurrent layer size	256
activation function	ReLU
weight decay	0.0001
parallel environments	200
episode length	400
epoch	500
chunk length	100
learning rate	0.0005
KL penalty coefficient β	$0 \rightarrow 1$
latent variable dimension	16

Table 2: Hyperparameters for VAE models

F Human Evaluation

Since diversity of humans is a key component in our approach, in this section we report the demographics of our participants. The study demographics include a population of 54% female and 46% male participants. The user demographics skewed towards younger ages, with 39% of participants between ages $18-26$, 44% of participants between $27-37$, 11% of participants between $38-48$, and 6% of participants ages 49 and above. The game experiences of the participants are shown in Figure 4.

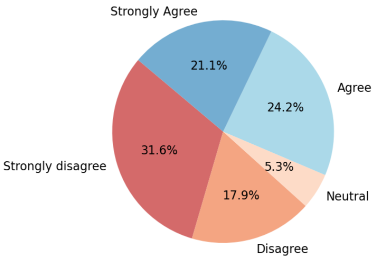


Figure 4: Percent of participants who agree with the statement “I have experience playing the game Overcooked”.

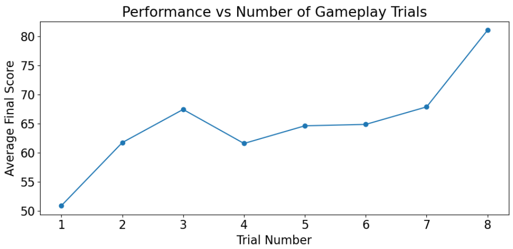


Figure 5: Human performance improves with the number of trials, indicating that the humans learn, change, and adapt their gameplay during the course of the evaluation.

F.1 Humans are adapting during evaluation

As shown by Fig 5, human performance improves with the number of trials, indicating that the humans learn, change, and adapt their gameplay during the course of the evaluation. In our evaluation with real humans, each user can change their strategy at any point in the game. A significant proportion of our users self-identify as novice players, both for video games and for Overcooked (17.9% and 49.5%, respectively, Figure 5, thus often exhibiting improved performance over the course of an increased number of trials. Figure 5 provides evidence of this pattern, demonstrating that humans change their gameplay style over the course of the evaluation.

G Additional Human Study Results

G.1 Human-AI team scores

Agent	Training data source	Counter circuit	Multi-strategy Counter
FCP	FCP-generated population	26.24 ± 0.96	34.35 ± 1.47
FCP + GAMMA		**57.87 ± 1.35**	**65.36 ± 1.24**
CoMeDi	CoMeDi-generated population	56.87 ± 1.43	27.11 ± 1.46
CoMeDi + GAMMA		**72.17 ± 1.61**	**34.72 ± 1.65**
MEP	MEP-generated population	76.19 ± 1.89	64.02 ± 2.07
MEP + GAMMA		**81.10 ± 2.03**	**88.30 ± 2.51**
PPO + BC	human data	53.51 ± 1.37	**85.26 ± 2.28**
PPO + BC + GAMMA		**59.67 ± 1.35**	77.53 ± 2.00
GAMMA HA	MEP pop. + human data	**91.11 ± 2.96**	**93.09 ± 3.19**

Table 3: Human evaluation results. Our methods (GAMMA) show significant improvements.

We conducted statistical significance tests and computed the p-value using the Holm-Bonferroni correction. See Table 4.

Hypothesis	Counter circuit (p-value)	Multi-(p-value)
FCP + GAMMA $>$ FCP	1.27×10^{-69}	1.43×10^{-50}
CoMeDi + GAMMA $>$ CoMeDi	7.31×10^{-12}	3.25×10^{-3}
MEP + GAMMA $>$ MEP	0.639	2.04×10^{-12}
PPO + BC + GAMMA $>$ PPO + BC	9.80×10^{-3}	7.48×10^{-2}
GAMMA HA $>$ FCP	4.21×10^{-113}	1.10×10^{-63}
GAMMA HA $>$ CoMeDi	1.55×10^{-23}	3.43×10^{-77}
GAMMA HA $>$ MEP	1.43×10^{-4}	2.92×10^{-13}
GAMMA HA $>$ PPO + BC	3.14×10^{-32}	0.426

Table 4: Statistical significance ($p < 0.05$) is achieved for the majority of the results.

G.2 Qualitative analysis

We include additional analyses of users self-reported responses for qualitative questions from our user study. Figure 6 shows the results for users' response to the agents' ability to adapt. The results indicate that two of our methods, **FCP + GAMMA** and **GAMMA-HA-DFT**, consistently receive higher ratings indicating better ability to adapt than their respective baseline agents, across both layouts. Figure 7 displays users' ratings for whether the agents demonstrated human-like behavior. The responses show that **FCP + GAMMA**, **CoMeDi + GAMMA**, and **GAMMA-HA-DFT** exhibit the most human-like gameplay in both layouts. Figure 8 includes responses for whether the agents' behavior was frustrating. Individuals consistently reported that **FCP + GAMMA** and **GAMMA-HA-DFT** demonstrated the least frustrating behavior.

Furthermore, we provide additional participant feedback for our agents from the user study as follows, starting with feedback for agents using our methods:

Feedback for **FCP + GAMMA**:

- "It was very coordinated."
- "This agent figured things out the fastest and worked with me well."
- "When it was in my way, it moved out of the way instead of blocking the path, which was an issue with other agents."
- "Much better cooperation by comparison."

Feedback for **CoMeDi + GAMMA**:

- "He was the best teammate."
- "This one did well, it moved around me enough to complete tasks and we were relatively efficient passing items around."
- "The agent figured out the next step I would have done."

Feedback for **PPO + BC + GAMMA**:

- "It was very efficient and placed things down so I could grab it on the other side."
- "This agent was very collaborative. All I did was put onions out and he did the rest of the work - super smooth."
- "It behaved smoothly and intentionally. It responded to my moves and helped rather than getting in the way."

In contrast, the feedback for the baseline agents is overall more negative, including themes such as frustration, inefficiency, and lack of cooperation.

Feedback for **FCP** (baseline):

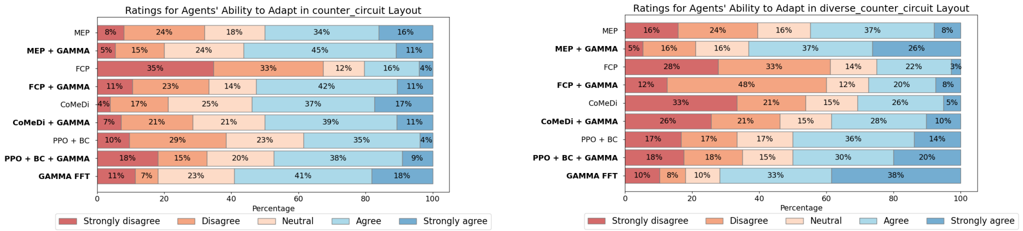
- "Frustrating."
- "He got in the way and didn't help at all."
- "Just seemed to lack coordination and felt less intelligent."

Feedback for **CoMeDi** (baseline):

- "The agent was helpful, but still did not fully cooperate as well as I thought it should have."
- "The agent was executing some random move actions before actually cooperating."

Feedback for **PPO + BC** (baseline):

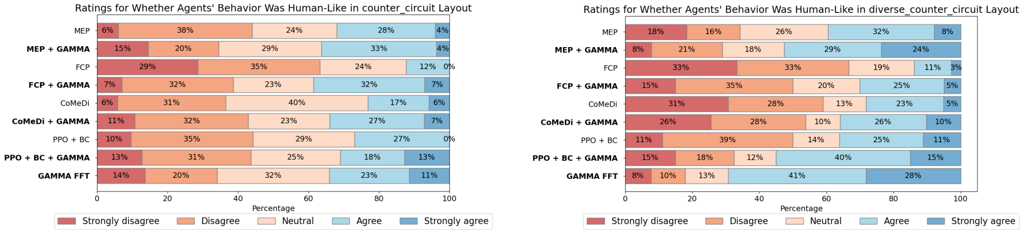
- "The agent was not helping much, it was doing its own thing, which meant we were not coordinating well."
- "Agent did everything by himself; didn't see the onions on the counter that I had placed; got in my way a bunch."
- "Very inefficient and lacking intelligence."



(a) Counter circuit layout

(b) Multi-strategy Counter

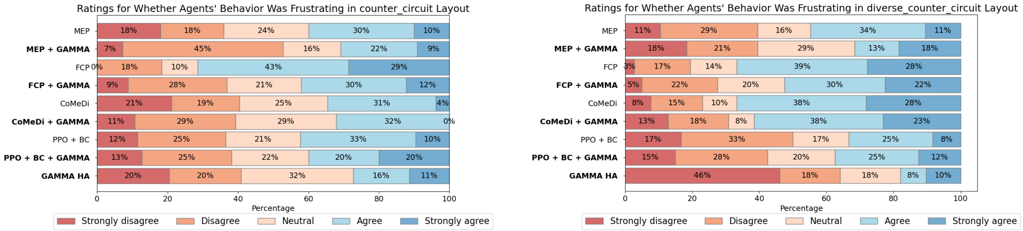
Figure 6: Human ratings for different agents. Individuals were asked to respond to the following question: "The agent adapted to me when making decisions: {Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree}". **FCP + GAMMA** and **GAMMA-HA-DFT** consistently receive higher ratings for ability to adapt compared to their respective baselines.



(a) Counter circuit layout

(b) Multi-strategy Counter

Figure 7: Human ratings for different agents. Individuals were asked to respond to the following question: "The agent's actions were human-like: {Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree}". **FCP + GAMMA**, **CoMeDi + GAMMA**, and **GAMMA-HA-DFT** consistently receive ratings for more human-like behavior compared to their respective baselines.



(a) Counter circuit layout

(b) Multi-strategy Counter

Figure 8: Human ratings for different agents. Individuals were asked to respond to the following question: "The agent's behavior was frustrating: {Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree}". **FCP + GAMMA** and **GAMMA-HA** consistently receive ratings for less frustrating behavior compared to their respective baselines.

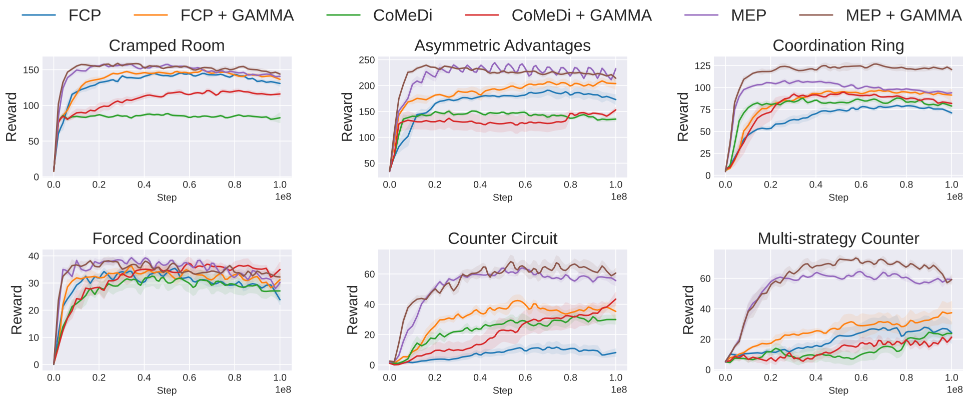
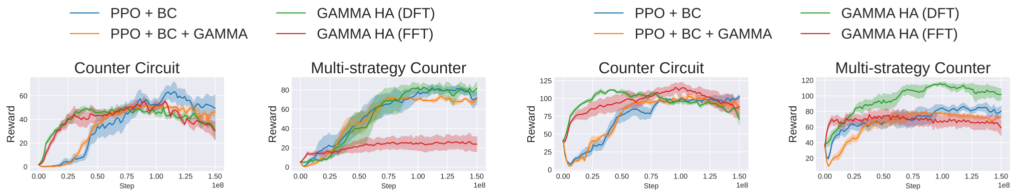


Figure 9: Learning curves for methods using simulated data across six layouts. Error bars are the Standard Error of the Mean (SE). All methods are evaluated using a held-out human proxy model as the partner player. **GAMMA** consistently shows better or equal performance on all layouts for both simulated data-generation methods (FCP, CoMeDi, MEP) when evaluated against the human-proxy model.



(a) Evaluation against human proxy agent.

(b) Evaluation against held-out self-play agents.

Figure 10: Learning curves for methods using human data. When evaluated with a held-out human proxy agent (a), human adaptive sampling learns faster on *Counter Circuit*, but does not reach better final performance since PPO-BC is trained to exploit a human-proxy agent. With simulated self-play partners (b), Human-Adaptive **GAMMA** with DFT shows better performance.

H Additional simulated results

Figure 9 provides the original data for the performance on simulated data.

In addition to training the generative model on the human dataset (**PPO-BC-GAMMA**, we can also leverage Human Adaptive (HA) sampling on the generative model: (**GAMMA-HA**). The learning curves of these methods and the baseline **PPO-BC** [1] are shown in Figure 10. When evaluated with a held-out human proxy agent, our methods do not show a great advantage. This is expected, because PPO-BC is trained to exploit a human proxy agent. However, in the human study, we show that this hurts its adaptation to more diverse real human players. On the other hand, we note that **GAMMA-HA** shows much faster learning in both layouts. Figure 10b shows that **GAMMA-HA** also adapts better to held-out self-play agents compared to PPO-BC.

I Compare decoder-only fine-tuning and full fine-tuning

I.1 Full fine-tuning can suffer from insufficient human data

In some early experiments, we find that only fine-tuning the decoder with human data (**GAMMA-HA-DFT**) provides consistently strong performance on both layouts, whereas full fine-tuning (**GAMMA-HA-FFT**) provides the strongest performance on the first layout, but weaker performance on the second layout. At first, we hypothesize this is because the second layout is more complex, but the number of human coordination trajectories available for training ($N = 11$) is significantly less than the first layout ($N = 37$). With more diverse potential strategies and less human data, we find the

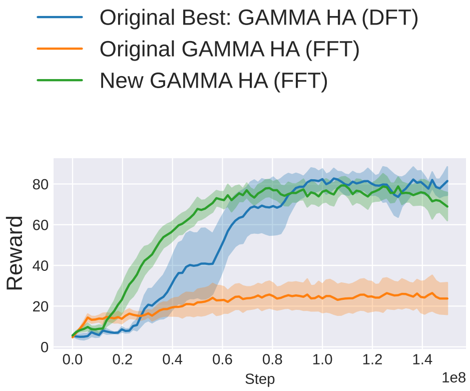


Figure 11: With larger KL Divergence penalty coefficient (β in Eq [1]), the new full-model fine-tuning (FFT) largely improves the original FFT and achieves a comparable performance with the original best decoder-only fine-tuning (DFT) method.

data is insufficient to fully fine-tune the generative model for the second layout. Since the entire model is fully fine-tuned, there is a higher chance that the model overfits the training human samples when the amount of human data is extremely small. Therefore, we believe the results of the FFT model could be improved were we to collect more data. However, it is realistic to test the scenario where limited human data is available, since human data can be quite expensive and difficult to collect. We find that in this low data regime, **GAMMA HA DFT** still provides excellent performance. Whether to fine-tune the encoder is a design choice that can be tuned for a particular domain based on the available data and the performance against simulated agents (since Figure 10 shows that FFT performed poorly here as well).

Agent	Training data source	Counter circuit	Multi-strategy Counter
FCP	FCP-generated population	32.11 ± 1.50	44.22 ± 2.15
FCP + GAMMA		**77.75** ± **2.09**	**74.44** ± **2.12**
CoMeDi	CoMeDi-generated population	69.62 ± 3.31	**32.02** ± **1.89**
CoMeDi + GAMMA		**77.77** ± **2.34**	**32.34** ± **1.79**
PPO + BC	human data	61.77 ± 2.53	97.73 ± 1.90
PPO + BC + GAMMA		72.32 ± 1.71	95.72 ± 1.75
GAMMA HA DFT	human data + FCP-generated population	82.82 ± 1.84	**103.76** ± **1.96**
GAMMA HA FFT		**91.84** ± **2.91**	34.05 ± 2.01

Table 5: Human evaluation results (outdated). Our methods (GAMMA) show significant improvements. However, GAMMA HA FFT is not stable on "Multi-strategy Counter".

I.2 Large regularization mitigates the problem of insufficient human data

Given the hypothesis that the human dataset is too small compared to the complexity of the *Multi-strategy Counter* environment, we find out that with a larger regularization over the fine-tuned model on the original model, we can mitigate this issue. See Figure 11.

J Can we condition the Cooperator policy on z?

One potential future work to improve the efficiency of online adaptation is to condition the policy on z. During testing with novel humans, the Cooperator can then infer the z of the human player and adapt to that z.

This method is orthogonal to our contributions where we focus on sampling partners with different z to train the Cooperator. We did some preliminary work to test the z-conditioned Cooperator. As shown in Figure 12, the performance of the z-conditioned Cooperator is not stable, it learns faster but the performance decreases when it is trained longer. Therefore, we do not include this method in

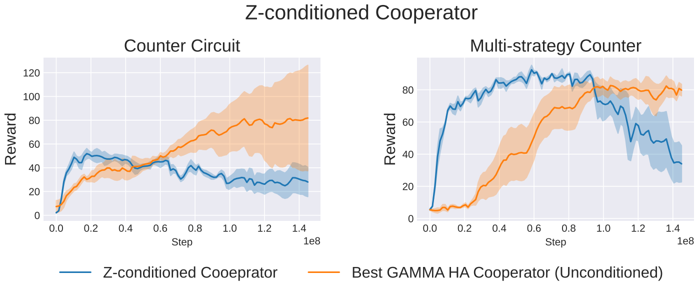


Figure 12: Performance of z-conditioned Cooperator. The z-conditioned Cooperator reaches a higher reward in the Multi-strategy Counter. The performance decreases after the peak since the z-conditioned policy overfits the encoder.

our main experiment, but future work about how to better train the $z-$ conditioned Cooperator is promising.