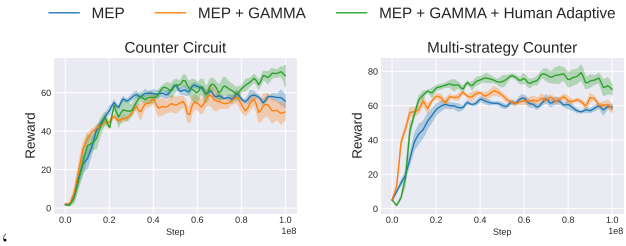
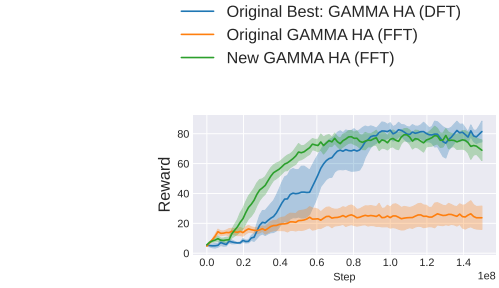
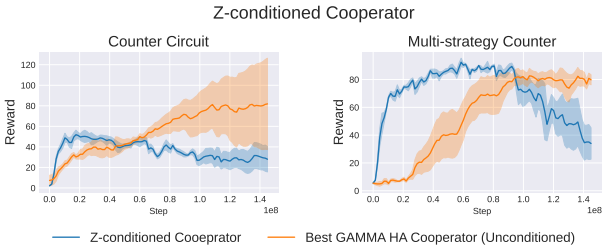


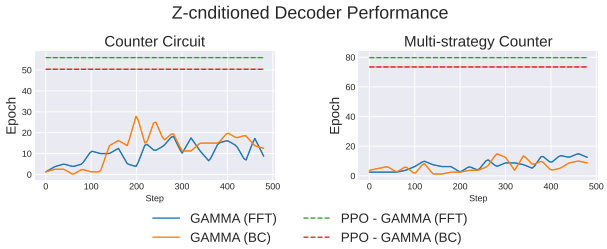
(a) **Results for a new baseline, MEP,** as requested by reviewer VM8E. GAMMA outperforms the baseline, and the gap is larger in the more complex Multi-strategy counter layout. The evaluation is done by playing against the BC human proxy model.



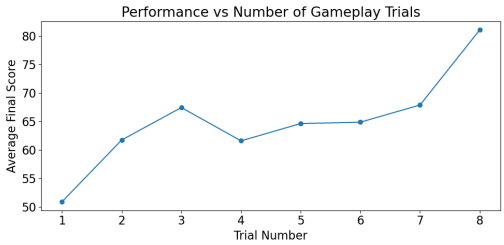
(b) With larger KL Divergence penalty coefficient (β in Eq [1]), the new full-model fine-tuning (FFT) largely improves the original FFT and achieves a comparable performance with the original best decoder-only fine-tuning (DFT) method. For reviewer UQLZ.



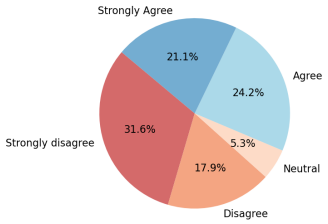
(c) Performance of z-conditioned Cooperator as requested by reviewer pvTn. The $z-$conditioned Cooperator reaches a higher reward in the Multi-strategy Counter. The performance decreases after peak since the $z-$conditioned policy overfits to the encoder



(d) Performance of z-conditioned decoders and unconditioned Cooperator against those decoders. Here GAMMA (FFT) is the fine-tuned VAE model and GAMMA (BC) the the VAE model trained solely on human data. Requested by pvTn.



(e) 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) Percent of participants who agree with statement "I have experience playing the game Overcooked"

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 1: Human evaluation results. Our methods (GAMMA) show statistically significant improvements.