Egocentric Team AI: Enabling Tactical Reasoning from the Operator's View

Soham Hans[1,2], **Yunzhe Wang**[1,2], **Volkan Ustun**[1]

[1]USC Institute for Creative Technologies
[2]University of Southern California
sohamhan@usc.edu, yunzhewa@usc.edu. ustun@ict.usc.edu

Abstract

The future of human–AI integration in high-stakes environments—ranging from defense operations to emergency response—requires AI systems that function as intuitive teammates rather than isolated tools. Yet a persistent perspective gap remains: video understanding models rely on third-person, broadcast-style views, while multi-agent reinforcement learning (MARL) systems operate on egocentric inputs but often depend on centralized critics during training, reducing the need for decentralized policies to internalize team-centric structure. As a result, agents may optimize behavior without learning policy-level representations of the tactical picture grounded in their own observations. To address this gap, we propose Egocentric Team AI, a research direction centered on learning implicit, distributed Common Operating Pictures from first-person views. Building on our prior work with the X-Ego-CS dataset and Cross-Ego Contrastive Learning (CECL), we outline how cross-egocentric representation alignment can be extended from passive video understanding to active multi-agent control. Specifically, we propose integrating cross-ego contrastive objectives into MARL within a multi-agent Doom-based environment as a testbed for decentralized, team-aware policy learning. By positioning cross-egocentric alignment as an inductive bias for decentralized coordination, this work charts a path toward embodied systems capable of adaptive cooperation without explicit communication.

Introduction

In high-stakes tactical domains, effective teamwork is rarely the result of constant verbal communication. Instead, it relies on a shared mental model of the evolving environment. Consider a mechanized infantry platoon executing an ingress into a hostile urban sector. As the lead Bradley Fighting Vehicle pivots its turret to scan a high-rise, the wingmen in the formation instinctively adjust their sectors of fire to cover the exposed flank. Simultaneously, a dismounted fire team performs a "stack" to clear a building: the point man focuses on the immediate threshold, while the rear security guard monitors the hallway behind. No words are spoken.

In command centers, this coordination is maintained via a Common Operating Picture (COP)—a centralized display of troops and hazards (Endsley and Jones 2025; Alberts et al. 2001). However, for the operator on the ground, the COP is not a static screen but a dynamic cognitive construction. To coordinate effectively, a soldier must transcend their immediate egocentric (first-person) sensory limits to mentally build an implicit COP of the broader allocentric state. They must utilize their limited view to infer the full tactical picture, intuitively understanding: "If my wingman is suppressing left, I must cover right."

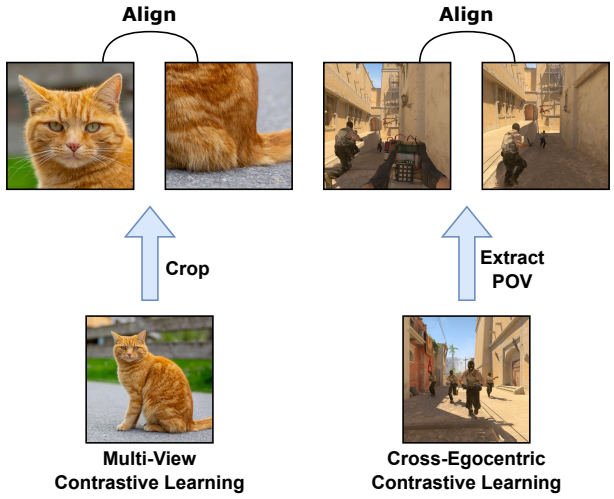

Figure 1: Illustration of Cross-Ego Contrastive Learning (CECL), where teammates' egocentric video representations are aligned across teams.

Replicating this level of tactical reasoning in artificial intelligence remains a formidable challenge due to a persistent perspective gap in current research. Current approaches to visual decision-making tend to fall into two extremes, neither of which fully captures the realities of tactical teamwork.

The Broadcast Perspective. In video understanding and sports analytics (e.g., SoccerNet, TacticAI (Deliege et al. 2021; Wang et al. 2024)), models rely on third-person, "God's eye" views. While effective for analyzing global formations, these approaches fail to model the partial observ-

ability and sensory constraints faced by embedded operators.

The Isolated Egocentric Perspective. Conversely, recent generalist embodied agents (e.g., SIMA, RT-2 (Bolton et al. 2025; Brohan et al. 2024)) and standard Multi-Agent Reinforcement Learning (MARL) approaches operate on first-person inputs but often lack mechanisms to intrinsically model team dynamics. Under centralized training, decentralized execution (CTDE), agents may rely on centralized critics with access to global state during training (Grönauer and Diepold 2022; Kharrat et al. 2025), reducing the pressure for decentralized policies to internalize team-centric structure within their own representations. As a result, policies may optimize behavior without explicitly encoding an implicit, distributed understanding of team state from egocentric observations alone (Hernandez-Leal, Kartal, and Taylor 2019).

We argue that enabling true AI teammates—whether autonomous robotic swarms or virtual wingmen—requires bridging this divide. Agents must learn to construct an implicit, distributed Common Operating Picture directly from egocentric inputs, such that decentralized policies encode team-relevant structure without relying on privileged global information.

The growing availability of structured esports datasets for *Counter-Strike* gameplay, such as ESTA (Xenopoulos and Silva 2022), has enabled large-scale analysis of professional matches; however, these resources lack synchronized cross-egocentric video streams across all players. Leveraging this ecosystem, we introduced X-Ego-CS in our prior work (Wang, Hans, and Ustun 2025), the first video action dataset for cross-egocentric multi-agent understanding in professional *Counter-Strike 2*. Using this dataset, we developed Cross-Ego Contrastive Learning (CECL), a representation learning framework that aligns teammates' egocentric views to induce shared latent team state (Figure 1). We showed that CECL improves prediction of team, enemy, and global states from egocentric video, suggesting that cross-egocentric alignment can encode implicit COP-like structure.

However, passive prediction is only a precursor to decision-making. The broader research challenge is to determine whether such cross-egocentric representations can serve as an inductive bias for decentralized policy learning in cooperative control settings.

In this paper, we articulate a research direction that extends CECL from passive video understanding to active multi-agent reinforcement learning. We propose integrating cross-ego contrastive objectives into MARL within a visually grounded, multi-agent Doom environment as a controlled testbed for studying decentralized team reasoning under partial observability. Rather than presenting conclusive empirical claims, we outline how such integration may enable policies to internalize team-level structure directly within egocentric representations.

By positioning cross-egocentric alignment as a coordination prior for MARL, this work charts a path toward embodied, team-aware systems capable of adaptive cooperation without explicit communication. Our aim is to establish conceptual foundations and experimental directions for Egocentric Team AI, moving toward agents that can see, think, and decide from the operator's view.

Contributions. This work advances a research direction toward Egocentric Team AI through three primary contributions:

- **Egocentric Team AI formulation.** We articulate team-centric tactical reasoning as the problem of constructing an implicit, distributed Common Operating Picture from egocentric observations, addressing the perspective gap between broadcast-view analysis and isolated egocentric decision-making.
- **Cross-ego alignment as a coordination prior.** We extend Cross-Ego Contrastive Learning (CECL) from passive video analysis to active multi-agent settings, positioning cross-perspective representation alignment as an inductive bias for learning decentralized, team-aware policy representations.
- **Toward decentralized team reasoning in MARL.** We outline how cross-ego contrastive objectives can be integrated into Multi-Agent Reinforcement Learning within a multi-agent Doom environment as a testbed for studying decentralized coordination under partial observability.

X-Ego-CS Dataset

The X-Ego-CS dataset provides the empirical foundation for the cross-egocentric representation learning framework used in this paper (Wang, Hans, and Ustun 2025). X-Ego-CS was designed to support the study of team-centric situational understanding from egocentric observations by capturing synchronized first-person views and structured interaction data from professional-level gameplay.

The dataset consists of highly curated gameplay recordings extracted from in-game replay demo files. Each match includes synchronized cross-egocentric video streams from all players, along with structured state–action trajectories derived from mouse and keyboard inputs and round-level event annotations. This combination enables joint analysis of perception, action, and team coordination under partial observability.

X-Ego-CS contains 45 professional matches spanning 1,011 rounds and approximately 124 hours of gameplay footage, covering 372 unique players. The average round duration is 44 seconds of player-alive time. All video streams are recorded at a resolution of 720p and 30 frames per second. To the best of our knowledge, X-Ego-CS is the first dataset to provide synchronized egocentric video streams and structured state–action trajectories for all players in professional e-sports matches. X-EGO-CS is available at: https://huggingface.co/datasets/wangyz1999/X-EGO-CS.

Cross-Egocentric Contrastive Learning

We briefly summarize Cross-Egocentric Contrastive Learning (CECL), introduced in our prior work (Wang, Hans, and

Perspective	Avg Δ	Best Task	Best Δ
Global	↑7.38%	global_roundWinner	↑23.25%
Enemy	↑4.60%	enemy_aliveCount	↑21.47%
Teammate	↑5.59%	teammate_aliveCount	↑28.00%
Self	↓6.49%	self_kill_20s	↑6.38%

↑ / ↓ indicate relative improvement or degradation.

Table 1: Categorical summary of CECL's impact across perspectives. Improvement is measured as relative percentage change in accuracy-based metrics, aggregated across the encoder models used in the experiments.

Ustun 2025), which serves as the representational foundation for the approach explored in this paper. CECL addresses the problem of learning team-centric visual representations from egocentric observations by aligning synchronized first-person views across teammates.

Formally, we consider egocentric video segments $\mathcal{V} \in \mathbb{R}^{A\times T\times 3\times H\times W}$, where A denotes the number of agents and T the number of frames per segment. Each agent's observation V_i is encoded by a vision backbone $f(\cdot)$ and projected into a shared embedding space via a linear head $g(\cdot)$, producing L2-normalized representations

$$\mathbf{u}_i = \frac{g(f(V_i))}{\|g(f(V_i))\|_2}.$$

CECL employs a multi-positive contrastive objective in which all teammates observing the same round at the same time segment are treated as positive pairs, while all other agent pairs serve as negatives. This formulation enforces alignment across divergent egocentric viewpoints and encourages the encoder to capture latent factors that are shared at the team level—such as formation, timing, and tactical phase—rather than viewpoint-specific visual details. As a result, the learned representation supports the construction of an implicit, distributed Common Operating Picture from partial observations.

Following (Zhai et al. 2023), CECL is optimized using a sigmoid-based contrastive loss:

$$\mathcal{L}_{\text{CECL}} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\log\frac{1}{1+\exp(m_{ij}(-t\,\mathbf{u}_i\cdot\mathbf{u}_j+b))}, \tag{1}$$

where $m_{ij} = 1$ for synchronized teammates and -1 otherwise, $\mathbf{u}_i \cdot \mathbf{u}_j$ denotes cosine similarity, and t and b are learnable temperature and bias terms. The bias is initialized to reflect the prior imbalance between positive and negative pairs, as detailed in our earlier work.

Intuitively, CECL aligns egocentric observations that are temporally and tactically correlated, such as synchronized sensory disruptions or coordinated movement patterns, forcing the encoder to map these shared cues to a common latent representation. While CECL was originally evaluated in passive video understanding settings, in this paper we treat it as a fixed representation learning mechanism and study how such cross-egocentric alignment can be leveraged in active multi-agent decision-making. In our prior work, by treating X-Ego-CS as a fixed data source and leveraging its cross-egocentric structure to study representation learning and coordination, we demonstrated improvements in an agent's ability to predict team, enemy, and global states from video as summarized in Table 1.

Method

Cooperative Multi-Agent Learning Setup

To investigate Egocentric Team AI in an active control setting, we consider cooperative multi-agent reinforcement learning under partial observability in visually grounded environments (Figure 2). Multiple agents share a common task and reward function but receive distinct egocentric, first-person visual observations. Each agent must therefore act solely from its own limited view while implicitly reasoning about the broader team state—analogous to constructing a decentralized Common Operating Picture.

This setting exposes the perspective gap described earlier: although execution is decentralized and egocentric, standard MARL training often relies on centralized signals that reduce the need for policies to internalize team-level structure. Without additional inductive biases, this can produce representational inconsistencies that hinder coordination.

As a research testbed, we focus on centralized training, decentralized execution (CTDE) frameworks such as Multi-Agent Proximal Policy Optimization (MAPPO) (Yu et al. 2022) and Multi-Agent Soft Actor-Critic (MASAC) (Bettini, Prorok, and Moens 2024). In these frameworks, decentralized policies operate on local observations at execution time, while a centralized critic aggregates multi-agent information during training. Our goal is not to propose a new MARL algorithm, but to examine how cross-egocentric representation alignment can be integrated into existing CTDE methods so that decentralized policies more effectively encode team-centric structure from egocentric inputs.

Visual Representations and Cross-Egocentric Contrastive Learning

Within this CTDE setting, each agent's recent egocentric observations are encoded into a fixed-dimensional embedding by a vision backbone. In a baseline configuration, these embeddings are passed directly to the actor (and centralized critic during training) to produce action distributions. Under this setup, agents learn independently from their own egocentric representations, without any mechanism to align representations across teammates.

To address this limitation, we augment the pipeline with Cross-Egocentric Contrastive Learning (CECL) as a representation-level coordination prior. A trainable projection head is inserted between the vision encoder and the policy and critic networks (Figure 2). Embeddings from different agents at the same timestep are treated as positive pairs and aligned via a contrastive objective, while embeddings from other timesteps or episodes serve as negatives.

The contrastive objective is optimized jointly with the MARL loss. Conceptually, this auxiliary signal encourages agents to map distinct egocentric views into a shared latent

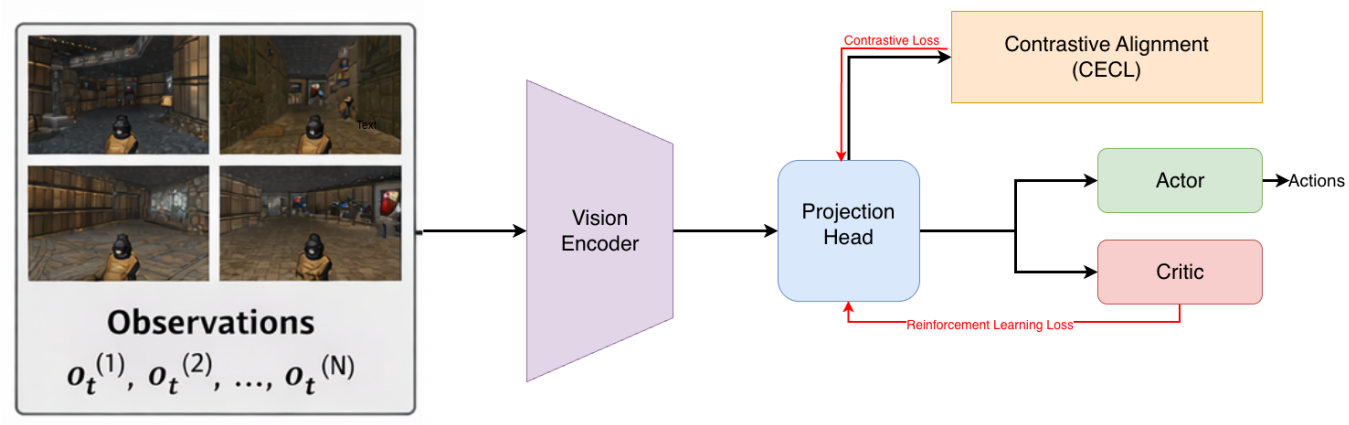

Figure 2: Cross-Egocentric Contrastive Learning for Cooperative Multi-Agent Control.

space, reducing representational mismatch and promoting decentralized, team-aware policy representations.

Proposed Evaluation Testbed

To study this integration in a controlled setting, we plan to evaluate CECL-augmented MARL within a visually grounded multi-agent environment based on the DOOM game engine, using the VizDoom framework (Kempka et al. 2016). VizDoom provides high-dimensional, first-person visual observations and has been widely used for perception-driven decision making under partial observability. Its first-person perspective and fast-paced dynamics make it a suitable testbed for examining whether cross-egocentric alignment can facilitate decentralized coordination from egocentric visual input.

Research Agenda and Evaluation Framework

We outline a research agenda for advancing Egocentric Team AI through the integration of cross-egocentric representation learning within cooperative multi-agent reinforcement learning.

Core Research Questions

The central hypothesis underlying this direction is that aligning egocentric representations across teammates can induce decentralized policies that internalize team-centric structure. To evaluate this hypothesis, we focus on the following questions:

- **Representation Internalization:** Does cross-egocentric alignment encourage decentralized policies to encode implicit team state without reliance on centralized critics at execution time?
- **Coordination Efficiency:** Can representation-level alignment improve decentralized coordination under partial observability?
- **Generalization:** Do aligned representations transfer across task variations and environment configurations?

Evaluation Metrics

To study these questions, evaluation should extend beyond aggregate reward. We propose two complementary categories of metrics:

- **Coordination Metrics:** joint survival time, task completion efficiency, spatial coverage, formation stability, and other measures of collective performance that reflect effective decentralized cooperation.
- **Representation Diagnostics:** linear probes on latent embeddings to predict teammate and opponent state, measuring the degree to which decentralized representations encode implicit Common Operating Picture structure.

Together, these metrics provide a principled framework for assessing whether cross-egocentric alignment meaningfully shapes policy-level representations and supports decentralized team reasoning.

Conclusion

Egocentric Team AI offers a research direction for closing the perspective gap between broadcast-style analysis and decentralized egocentric decision-making. We argue that effective AI teammates must construct an implicit, distributed Common Operating Picture directly from first-person observations, rather than relying on centralized training signals or privileged global state.

Building on X-Ego-CS and Cross-Ego Contrastive Learning (CECL), we position cross-egocentric representation alignment as an inductive bias for decentralized coordination in multi-agent reinforcement learning. We outline how integrating such alignment within CTDE frameworks and evaluating it in a multi-agent Doom testbed can advance the study of team-aware policy learning under partial observability.

By centering decision-making on the operator's view, this work charts a path toward embodied, communication-efficient systems capable of adaptive cooperation in high-stakes environments.

Acknowledgments

The authors acknowledge the use of Large Language Models for assistance with proofreading and grammar checking. All content was reviewed, edited, and approved by the human authors, who take full responsibility for the final manuscript. The project or effort depicted was or is sponsored by the U.S. Army Combat Capabilities Development Command – Soldier Centers under contract number W912CG-24-D-0001. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Alberts, D. S.; Garstka, J. J.; Hayes, R. E.; and Signori, D. A. 2001. Understanding information age warfare.

Bettini, M.; Prorok, A.; and Moens, V. 2024. Benchmarl: Benchmarking multi-agent reinforcement learning. *Journal of Machine Learning Research*, 25(217): 1–10.

Bolton, A.; Lerchner, A.; Cordell, A.; Moufarek, A.; Bolt, A.; Lampinen, A.; Mitenkova, A.; Hallingstad, A. O.; Vujatovic, B.; Li, B.; et al. 2025. Sima 2: A generalist embodied agent for virtual worlds. *arXiv preprint arXiv:2512.04797*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2024. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. *URL https://arxiv. org/abs/2307.15818*.

Deliege, A.; Cioppa, A.; Giancola, S.; Seikavandi, M. J.; Dueholm, J. V.; Nasrollahi, K.; Ghanem, B.; Moeslund, T. B.; and Van Droogenbroeck, M. 2021. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4508–4519.

Endsley, M. R.; and Jones, D. G. 2025. *Designing for situation awareness: An approach to user-centered design*. CRC press.

Grönauer, S.; and Diepold, K. 2022. Deep multi-agent reinforcement learning: a survey. *Artificial Intelligence Review*, 55: 895–943.

Hernandez-Leal, P.; Kartal, B.; and Taylor, M. E. 2019. A Survey and Critique of Multiagent Deep Reinforcement Learning. *Autonomous Agents and Multi-Agent Systems*.

Kempka, M.; Wydmuch, M.; Runc, G.; Toczek, J.; and Jaśkowski, W. 2016. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *2016 IEEE conference on computational intelligence and games (CIG)*, 1–8. IEEE.

Kharrat, S.; Fourati, F.; Canini, M.; Alouini, M.-S.; and Aggarwal, V. 2025. Latent Inference for Effective Multi-Agent Reinforcement Learning under Partial Observability. *European Workshop on Reinforcement Learning (EWRL)*.

Wang, Y.; Hans, S.; and Ustun, V. 2025. X-Ego: Acquiring Team-Level Tactical Situational Awareness via Cross-Egocentric Contrastive Video Representation Learning. *arXiv preprint arXiv:2510.19150*.

Wang, Z.; Veličković, P.; Hennes, D.; Tomašev, N.; Prince, L.; Kaisers, M.; Bachrach, Y.; Elie, R.; Wenliang, L. K.; Piccinini, F.; et al. 2024. TacticAI: an AI assistant for football tactics. *Nature communications*, 15(1): 1906.

Xenopoulos, P.; and Silva, C. 2022. Esta: An esports trajectory and action dataset. *arXiv preprint arXiv:2209.09861*.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35: 24611–24624.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.