X-Ego: Acquiring Team-Level Tactical Situational Awareness via Cross-Egocentric Contrastive Video Representation Learning

Yunzhe Wang
University of Southern California
Los Angeles, United States
yunzhewa@usc.edu

Soham Hans
University of Southern California
Los Angeles, United States
sohamhan@usc.edu

Volkan Ustun
USC Institute for Creative Technologies
Los Angeles, United States
ustun@ict.usc.edu

ABSTRACT

Human team tactics emerge from each player's individual perspective and their ability to anticipate, interpret, and adapt to teammates's intentions. While advances in video understanding have improved the modeling of team interactions in sports, most existing work relies on third-person broadcast views and overlooks the synchronous, egocentric nature of multi-agent learning. We introduce **X-Ego-CS**, a benchmark dataset consisting of 124 hours of gameplay footage from 45 professional-level matches of the popular e-sports game *Counter-Strike 2*, designed to facilitate research on multi-agent decision-making in complex 3D environments. X-Ego-CS provides **cross-egocentric** video streams that synchronously capture all players' first-person perspectives, along with state-action trajectories. Building on this resource, we propose **Cross-Ego Contrastive Learning (CECL)**, which aligns teammates' egocentric visual streams to foster team-level tactical situational awareness from an individual's perspective. We evaluate CECL on a teammate-opponent location prediction task, demonstrating its effectiveness in enhancing agent's ability to infer both teammate and opponent positions from a single first-person view on state-of-the-art video encoders. Together, X-Ego-CS and CECL establish a foundation for cross-egocentric multi-agent benchmarking in esports. More broadly, our work positions gameplay understanding as a testbed for multi-agent modeling and tactical learning, with implications for spatiotemporal reasoning and human-AI teaming in both virtual and real-world domains. Code and dataset are available at https://github.com/HATS-ICT/x-ego

KEYWORDS

Contrastive Learning, Video Understanding, Gameplay Understanding, Teammate Modeling, Multi-Agent Systems, Esports Analytics

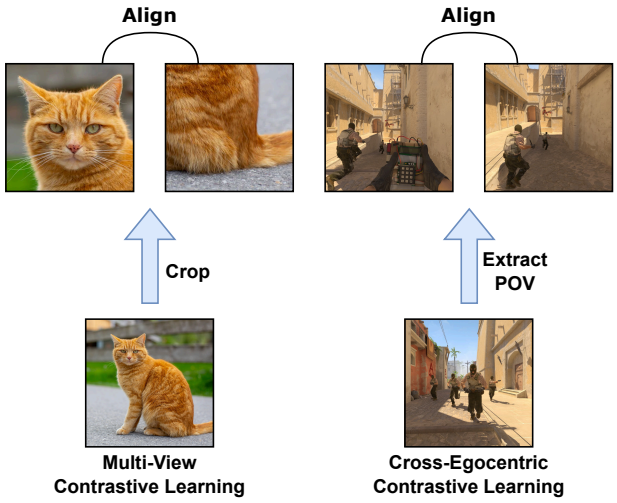


Figure 1: Illustration of Cross-Ego Contrastive Learning (CECL), where teammates' egocentric video representations are aligned across teams.

1 INTRODUCTION

Team tactics are a defining feature of many human activities, from competitive sports to defense operations, emergency response, and cooperative games. Understanding and modeling such tactics requires reasoning not only about high-level joint actions and spatiotemporal coordination, but also about each agent's capacity to infer and anticipate the beliefs, intentions, and reactions of both teammates and opponents—a capability often referred to as Computational Theory of Mind [4, 18, 24, 41]. Early computational approaches to teamwork primarily focused on learning joint policies within simulated Multi-Agent Reinforcement Learning (MARL) frameworks [5, 28, 32, 45, 52, 57]. However, these methods face two major limitations: (1) many simulation environments are grid-based or two-dimensional, simplifying real-world complexity; even in large-scale domains such as *StarCraft II* [52] and *Dota 2* [5], the worlds are presented through isometric or "pseudo-3D" perspectives rather than fully three-dimensional environments; (2) learning architectures seldom include explicit models of other agents' internal states. Consequently, while such systems can coordinate actions, they lack the ability to interpret or anticipate teammates' and opponents' mental states. This gap limits their scalability and transferability to real-world scenarios, which is essential for advancing the computational modeling of human team tactics.

In recent years, there has been growing interest in extending team-level behavior modeling to real-world competitive sports domains, particularly in soccer and basketball [9, 10, 17, 44, 55] for tasks such as action recognition, event detection, tactical planning and discovery, and automated commentary generation. This trend is driven both by the complexity of these sports and by the economic and analytical value of understanding team strategies and player interactions. However, existing datasets and models still fall short of enabling true multi-agent understanding since most datasets rely

Preprint. Under review.

on third-person broadcast footage from fixed angles with global visibility, which limits the study of individual perception, uncertainty, and coordination from a first-person perspective.

In contrast, First-Person Shooter (FPS) competitive video games such as *Counter-Strike* provide a natural testbed and an effective middle ground that balances game-state richness and decision-making complexity, while offering reliable ground-truth data for studying team behavior under partial observability. The game provides a cooperative yet adversarial environment on fixed 3D maps, where players must continuously balance tactical risk, spatiotemporal dynamics, and coordination. While relatively underexplored, prior work has begun to investigate this space: [56] introduced the ESTA dataset along with the AWPY parser for extracting ground-truth state-action trajectories and game events, [54] developed the DECOY simulator to support multi-agent reinforcement learning, [11] trained a transformer-based control model to mimic professional player movement, and [13] studies players' emotional dynamics with synchronized physiological, behavioral, and gameplay signals. While these efforts highlight the promise of the domain, they largely focus on individual agents or global game states, lacking a systematic approach to modeling the synchronous, egocentric nature of multi-agent team interactions.

To address this gap, we introduce **X-Ego-CS**, a benchmark dataset comprising 124 hours of gameplay footage from 45 professional-level *Counter-Strike 2* matches, containing synchronized egocentric video streams from all players. Unlike prior datasets focused on single perspectives or third-person views, X-Ego-CS captures simultaneous first-person recordings paired with state-action trajectories sampled at 64 ticks per second. We define this multi-perspective video paradigm as **cross-egocentric**, combining the egocentric focus of individual perception with cross-view synchronization across teammates.

Building upon X-Ego-CS, we propose **Cross-Ego Contrastive Learning (CECL)**, a method designed to learn shared world-state representations by aligning teammates' egocentric visual streams at corresponding timesteps. This alignment encourages models to internalize a collective understanding of team context, which is an essential capability for agents aiming to exhibit *theory-of-mind*-like reasoning and collaborative tactical awareness. We evaluate CECL on a downstream **teammate-opponent location prediction** task, demonstrating its effectiveness in enhancing agents' ability to infer both teammate and opponent positions from a single first-person view when applied to state-of-the-art video encoders. These results highlight CECL's potential to foster team-level situational awareness and establish **cross-egocentric multi-agent video understanding** as a promising direction for advancing tactical reasoning and human-AI teaming. Our contributions are twofold:

- We introduce X-Ego-CS, the first benchmark dataset for cross-egocentric multi-agent video understanding in esports, and established teammate-opponent location prediction task for teammate and opponent modeling.
- We propose CECL, a visual contrastive learning method that learns shared team-state representations by aligning entire teams' first-person views at the same timestep, boosting individual agents' team-level situational awareness.

2 RELATED WORK

Sports Understanding. Computational analysis of team sports has traditionally focused on third-person broadcast views, particularly in soccer and basketball, addressing tasks such as action recognition [10, 14], tactical planning [55], and automated commentary [36, 43, 44]. While these approaches have advanced strategic analysis and performance analytics, they rely on omniscient overhead perspectives that do not capture the egocentric, partially observable nature of real-time decision-making that players experience.

Gameplay Understanding. Research in gameplay understanding within AI can be broadly organized into two categories: (1) enabling AI agents to play games autonomously, and (2) facilitating effective collaboration between humans and AI. Significant progress in autonomous gameplay has been achieved through deep reinforcement learning in single-agent settings, from early successes in Atari [37] to mastering board games [47, 48], as well as through multi-agent reinforcement learning in cooperative-competitive environments, including multiplayer online battle arena (MOBA) games [5, 52], soccer [25], and emergent multi-agent behaviors [3]. More recent advances have explored LLM-enabled gameplay [53] and world modeling approaches [6, 20, 21].

In parallel, human-AI collaboration has emerged as a critical area in games and robotics, with works exploring coordination challenges [7], human expectations of AI collaborators [61], dual-process architectures for real-time teaming [62], shared mental model alignment tools [16], and multimodal research platforms [60]. Recent work has also proposed interpretable behavioral task-space frameworks to evaluate human-AI alignment in large-scale multiplayer games [46]. Analysis of teammate preferences reveals that interpretability can be as important as performance, with humans often favoring rule-based agents over learned ones despite comparable outcomes [49].

Despite these advances, open challenges remain in scaling gameplay understanding to competitive esports and AAA-level games, particularly in reasoning over rich visual environments, fast-paced dynamics, and high degrees of freedom, and in transferring knowledge learned from gameplay to real-world human-AI teaming.

Contrastive Learning. Contrastive learning has emerged as a powerful paradigm for self-supervised representation learning. Its core idea—learning by distinguishing positive from negative pairs—can be traced back to early energy-based models and noise-contrastive estimation [19, 27]. In natural language processing, similar principles underpin word embedding methods such as Word2Vec [35], which train models to discriminate true word-context pairs from random ones. More recently, contrastive learning has achieved remarkable success in computer vision through frameworks like CPC [39], SimCLR [8], and MoCo [22], which leverage data augmentations and large negative sets to learn transferable visual representations. Building on these advances, contrastive learning has also proven effective in multimodal settings, where it enables alignment between different modalities such as vision and language [42, 59].

In the multi-agent learning domain, contrastive learning has gained increasing traction in recent years for addressing key challenges such as coordination, diversity, opponent modeling, and communication. Recent work explores how contrastive objectives

can disentangle agent roles and enhance cooperation by promoting behavioral heterogeneity and more effective credit assignment [23]. Other studies investigate how contrastive trajectory representations can encourage diversity among agents that share parameters, leading to richer exploration and more robust team strategies [29]. To improve adaptability in mixed cooperative-competitive settings, contrastive formulations have been applied to learn policy embeddings of teammates and opponents directly from single-agent observations, enabling faster response and generalization across tasks [33]. In addition, contrastive principles have been used to learn more effective communication protocols between agents, capturing shared environmental information and promoting symmetric message exchange in decentralized systems [30, 58]. However, these approaches have primarily been validated in simplified reinforcement learning environments, raising important questions about their applicability and scalability in more realistic, high-complexity multi-agent scenarios.

3 X-EGO-CS DATASET

Our dataset comprises highly curated gameplay recordings including cross-egocentric video streams and structured state-action trajectories, all extracted from in-game replay demo files. We describe the data curation process in Section 3.1, and provide summary statistics in Section 3.2. To our best knowledge, X-Ego-CS is the first dataset that contains synchronized egocentric video streams and structured state-action trajectories from all players in professional e-sports matches. A comparison with similar datasets is shown in Table 1.

Table 1: Characteristic comparison of X-Ego-CS with similar video datasets. Symbols: ✓ = available, ✗ = not available, △ = partially satisfied.

Dataset	Domain	Team	Expert	Video	Traj.	Ego-Cen.
ESTA [56]	ESports (CS:GO)	✓	✓	✗	✓	✗
AMuCS [13]	ESports (CS:GO)	△	✗	✓	✓	✓
SoccerNet [14]	Sports (Soccer)	✓	✓	✓	△	✗
SoccerReplay [43]	Sports (Soccer)	✓	✓	✓	✗	✗
Waymo Open [50]	Self-Driving Cars	✗	✗	✓	✓	✓
Ego4D [15]	Daily Activities	✗	✗	✓	✓	✓
X-Ego-CS (Ours)	ESports (CS2)	✓	✓	✓	✓	✓

3.1 Automated Data Collection and Curation

We collected professional-level Counter-Strike 2 data from in-game demo files (`.dem`), all downloaded from the popular CS2 matchmaking platform FACEIT [12]. Using FACEIT's official API and following website policies, we retrieved the public Elo rating leaderboard and match history for the top 100 players, collecting their most recent matches. From these replays, we extracted structured metadata, player state-action trajectories, and game events using off-the-shelf parsers including AWPY and demoparser2 [26, 56].

To ensure reproducible and standardized video data, we developed an automated in-game recording system built on top of the Counter-Strike 2 demo playback engine. The demo metadata provides exact tick ranges for every player's alive period within a round. By issuing console commands such as `demo_gototick` to jump to the start tick, `spec_player` to lock the camera onto the correct player, and `demo_resume`/`demo_pause` to control playback, we were able to deterministically replay only the segments corresponding to each player's active life. During playback, NVIDIA Shadowplay [38] was automatically triggered via hotkeys to record the alive duration of each player per round.

After filtering for quality, the final dataset contains 45 matches, all played on the `de_mirage` map, including synchronized egocentric video recordings from all players and state-action trajectories.

3.2 Statistics

X-Ego-CS contains 45 curated professional-level matches spanning 1011 rounds and approximately 124 hours of gameplay video footage. Each match includes cross-egocentric video streams, structured state-action trajectories, and round-level event annotations, covering 372 unique players, with an average round duration of 44 seconds of player alive time. All videos are recorded at 720p and 30 FPS. Additional statistics are shown in Figure 2.

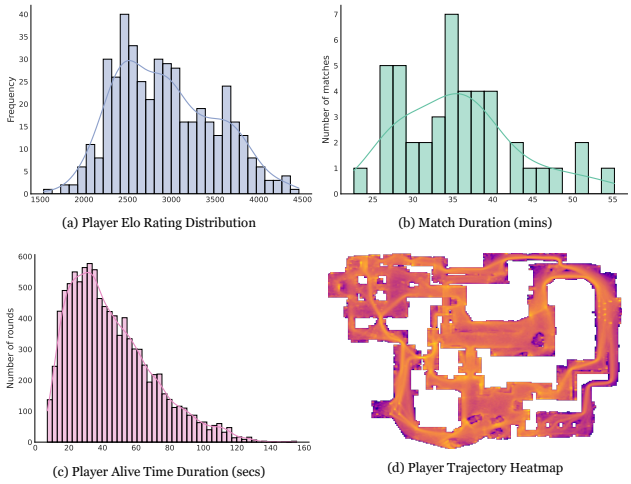


Figure 2: Additional statistics of the X-Ego-CS dataset. (a) Player FACEIT Elo Rating Distribution: above 2000 corresponds to the highest FACEIT level 10 rank (roughly top 10% of players), and above 3000 represents roughly the top 1%. (b) Match Duration. (c) Player Alive Time Duration, which corresponds to the duration of video files. (d) Player Trajectory Heatmap aggregated from both team sides across all matches, visualized on the de_mirage map.

4 METHOD

In this paper, we aim to learn a visual representation that enables individual agents to achieve team-level situational awareness from their egocentric observations. We first formulate cross-egocentric tactical understanding as a team-based representation learning task (Sec. 4.1), then describe our model architecture (Sec. 4.2) and Cross-Egocentric Contrastive Learning objective (Sec. 4.3), and finally present the downstream teammate-opponent location prediction tasks (Sec. 4.4).

4.1 Problem Formulation

We address the problem of multi-agent video understanding and teammate modeling with RGB video segments $\mathcal{V} \in \mathbb{R}^{A\times T\times 3\times H\times W}$ from each team, where A denoting the number of agents and T denoting the number of frames. Our goal is to learn a visual encoder Φ_{vision} that maps each agent's egocentric video segment to an embedding space $\mathbf{Z}$, such that $\text{Sim}(Z_i, Z_j)$ is maximized for teammates $i, j \in \mathcal{T}$ at the same time segment and minimized for agents otherwise. After alignment, we can take a single agent's embedding Z_i or a subset of agents from the same team to infer information about all teammates and opponents. We hypothesize that the contrastive objective encourages implicit information sharing among teammates by aligning their egocentric representations toward a shared latent team state. This process allows the encoder to capture common situational cues such as formation, timing, tactical phase.

Consider a flashbang grenade event that causes synchronous blindness among agents in the flash's line of sight. When multiple teammates are simultaneously flashed, their first-person visual streams exhibit similar characteristics (whiteout effects and reduced motion). The contrastive objective maximizes $Sim(Z_i, Z_j)$ for these affected teammates, which forces Φ_{vision} to produce similar embeddings for the shared visual patterns. Critically, the encoder cannot simply memorize individual whiteout frames—it must learn that such synchronized sensory disruptions encode tactical information: the grenade's origin and trajectory constrain opponent positions, and coordinated flashes typically indicate clustered team movements during site pushes. CECL trains the encoder to map these shared visual patterns to a latent representation of the underlying team state. Even when only one player's view is available during inference, the learned embedding space could help the model better infer teammate proximity and opponent positioning, as the alignment process and downsteam objectives encodes relevant spatial configurations correlated with such sensory patterns.

4.2 Model Architecture

Our model is built upon state-of-the-art vision transformer architecture (Sec 5.1) that processes egocentric video streams. The overall architecture consists of three main components: (1) a spatiotemporal video encoder that extracts frame-level and temporal features from each player's first-person view, (2) a contrastive projection head that maps these features into a shared embedding space where team-based alignment is enforced, and (3) downstream MLP predictor heads that predict either teammate or opponent locations from subset and aggregated agent embeddings (Fig 3).

A visual encoder-projector network Φ_{vision} extracts spatiotemporal representations for each agent, producing embeddings

$$E = \Phi(V) \in \mathbb{R}^{A\times d_{\text{proj}}}$$

During inference, a subset of N agents's visual embeddings is selected, where $N \le A$, and for the single-agent case $N = 1$. When $N > 1$, the selected embeddings are aggregated through an agent aggregator ϕ_{agg} with is concatenation plus two-layer MLP. The resulting representation is concatenated with the team-side embedding $S \in \mathbb{R}^{d_s}$ to form the combined feature

$$Z = [F; S]$$

where S is a single vector indicating the team side of the POV (either T team or CT team). This combined feature is then processed by task-specific heads

$$\Psi = \{\Psi_{\text{TLN}}, \Psi_{\text{ELN}}\}$$

corresponding to the **Teammate Location Nowcast (TLN)** and **Enemy Location Nowcast (ELN)** tasks. This formulation allows the unified model $\Psi(\Phi(V))$ to flexibly adapt to single-agent or multi-agent subsets for downstream objectives.

4.3 Cross-Egocentric Contrastive Learning

We adopt a sigmoid-based contrastive objective [59] to align cross-egocentric representations within teams. Given a video encoder $f(\cdot)$, for each player i in a batch, we obtain L2-normalized embeddings $\mathbf{u}_i = \frac{f(V_i)}{\|f(V_i)\|_2}$ from their egocentric video segment. We employ a *multi-positive contrastive strategy*, where all teammates observing the same round at the same time segment serve as positive pairs (Fig 3a). This naturally creates a three-level hierarchy of negatives: (1) same team and round but different time segments, (2) different team (i.e., opponents) at the same or different time, and (3) different rounds or matches. The contrastive objective is:

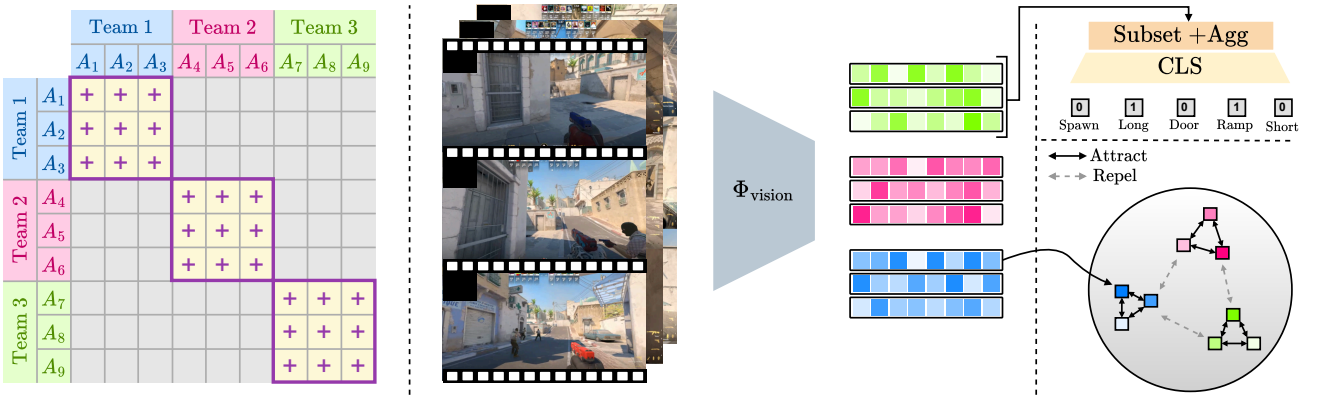
$$\mathcal{L}_{\text{CECL}} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\log\frac{1}{1+e^{m_{ij}(-t\,\mathbf{u}_i\cdot\mathbf{u}_j+b)}} \tag{1}$$

Here, $\mathbf{u}_i \cdot \mathbf{u}_j$ represents the cosine similarity between player i and player j's egocentric embeddings. The parameter t is a learnable temperature for scaling, and b is a learnable bias. The label m_{ij} equals 1 if players i and j are on the same team observing the same round at the same time segment (positives), and -1 otherwise (negatives). We initialize b to -3 and t to $\log(10)$. This initialization accounts for the large imbalance between positive and negative pairs where negatives dominate the loss, with a positive ratio of $p_+ = |\mathcal{B}|/|\mathcal{B}|^2 = 1/|\mathcal{B}|$. We initialize the bias to roughly the prior log-odds of the positive ratio, i.e., $\text{logit}(p_+) = \log(p_+/(1-p_+))$, which in our case is roughly -3. We also conducted an ablation of the bias terms in Section 5.3 that confirms the design choice.

4.4 Teammate-Opponent Location Prediction

We designed a video-based teammate-opponent location prediction task to evaluate the effectiveness of CECL. Given an agent's egocentric video segment, Teammate Location Nowcast (TLN) aims to predict the spatial positions of all teammates at the corresponding time step, while Opponent Location Nowcast (OLN) aims to infer the positions of all adversarial agents.

Formally, let $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$ denote the set of N discrete location identifiers in the environment. At time step t, given an agent's egocentric video segment $\mathbf{v}_t$, we formulate the location nowcasting task as a multi-label classification problem. For each location $l_i \in \mathcal{L}$, we predict a binary occupation label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates that at least one agent (teammate for TLN, opponent for OLN) currently occupies location l_i at time t, and $y_i = 0$ otherwise. The model outputs a prediction vector $\mathbf{y}_t = [y_1, y_2, \ldots, y_N]^T \in \{0, 1\}^N$, representing the concurrent occupancy distribution across all locations. In X-Ego-CS, on the de_mirage map, there are $N = 23$ locations.



(a) In-Batch Positives and Negatives (b) Cross-Egocentric Video Encoding (c) CECL and Downstream

Figure 3: Illustration of the forward pass on a data batch containing three teams of three agents each. (a) Agents from the same team are treated as positive pairs, while agents from different teams or time segments are treated as negatives. (b) Each agent's egocentric video segment is encoded into a representation vector; a mask is applied to the top-left corner to prevent leaking teammate or opponent locations from the mini-map. (c) The CECL objective aligns teammates' egocentric video representations within the same batch while repelling others. A subset of agents from a team is then selected for team-level location prediction, formulated as a multi-label classification task.

Classification Labels. We acquire location labels from the game trajectory data, where each 3D coordinate has a corresponding categorical area identifier. The `de_mirage` map contains 23 distinct locations with names corresponding to strategic callouts that players commonly use during gameplay, such as `T_Spawn`, `CT_Spawn`, `Bombsite_A`, `Bombsite_B`, `Catwalk`, `Stairs`, `Connector`, etc. To construct ground truth labels, we extract each agent's 3D coordinates at time t and map them to their categorical area name. When an agent's trajectory spans multiple areas within a video segment, we select the area at the middle temporal point to ensure alignment. For simplicity, we only use video segments where all 10 players are alive. The final ground truth $\mathbf{y}_t \in \{0,1\}^{23}$ is a binary vector where each dimension indicates whether at least one agent occupies the corresponding location.

Evaluation Metrics. We report standard metrics for multi-label classification: Subset Accuracy, which measures the fraction of samples where all location predictions exactly match the ground truth labels across all 23 locations; Hamming Distance, which quantifies the fraction of misclassified labels across all location predictions, and Micro and Macro F1 scores, averaged over all labels.

5 EXPERIMENTS

We evaluate CECL across multiple dimensions. We first describe the implementation details including encoder architectures, data processing, and training configurations (Sec. 5.1). We then present quantitative results comparing CECL against baseline models on single-POV and multi-POV settings (Sec. 5.2), followed by ablation studies on key hyperparameters (Sec. 5.3). Finally, we provide qualitative analysis through embedding space visualizations (Sec. 5.4).

5.1 Implementation Details

Video Encoders. We evaluate our cross-egocentric contrastive learning framework using four state-of-the-art video encoders: SigLIP [59], DINOv2 [40], ViViT [1], VideoMAE [51]. For image-based models (SigLIP, DINOv2), we compute temporal embeddings by mean pooling frame-level representations. For video models with fixed temporal receptive fields (ViViT, VideoMAE), we apply temporal resampling to match the target sequence length through interpolation or truncation.

All video encoders are kept frozen during training, and we learn a two-layer MLP projector $\phi_{\text{proj}} : \mathbb{R}^{d_{\text{enc}}} \to \mathbb{R}^{768}$ for contrastive alignment.

Data Processing. Videos are preprocessed through a standardized pipeline: (1) we resize videos to 224×224 resolution with bilinear interpolation, distorting the aspect ratio; (2) pixel values are normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$); and (3) temporal sampling at 4 FPS over 5-second windows with ± 0.3 second jittering for temporal robustness. The dataset is partitioned with a 70:15:15 train:validation:test split based on total rounds to ensure temporal independence. We mask out the top-left corner of all videos to ensure that teammate and opponent locations are not leaked in the mini-map.

Training Configuration. We employ the AdamW optimizer [31] with learning rate $\eta = 3 \times 10^{-4}$, weight decay $\lambda = 10^{-2}$, and bFloat16 data type with mixed-precision training. We train all models for 8 epochs and select the best checkpoint based on validation performance. We use batch size $B = 32$ and apply gradient accumulation when memory constraints require smaller effective batch sizes. All experiments are conducted on single GPUs including A100, A40, or RTX4090.

For the multi-label classification tasks Teammate Location Nowcast and Enemy Location Nowcast, we optimize the Binary Cross

Table 2: Performance comparison of CECL vs. baseline models on Teammate Location Nowcast and Enemy Location Nowcast tasks. Bold values indicate best results; red values highlight improvements over baseline.

Method	Teammate Location Nowcast				Enemy Location Nowcast			
	Sub.Acc ↑	Ham.Dist ↓	Macro F1 ↑	Micro F1 ↑	Sub.Acc ↑	Ham.Dist ↓	Macro F1 ↑	Micro F1 ↑
Off-the-shelf Models (Frozen) + Linear Probe								
DINOv2 [40]	16.89	9.90	35.66	57.19	13.47	**11.13**	**22.92**	**49.77**
ViViT [1]	16.97	10.33	35.69	56.12	**14.21**	11.30	22.83	48.14
VideoMAE [51]	12.92	11.46	21.00	47.25	12.57	11.50	18.10	45.64
SigLIP [59]	**17.67**	**9.58**	**39.77**	**59.40**	11.02	11.37	19.90	46.91
Cross-Ego Contrastive Learning								
CECL (DINOv2)	**18.13** (+1.24)	**9.67** (-0.23)	**37.61** (+1.95)	**58.37** (+1.18)	**14.60** (+1.13)	**11.00** (-0.13)	**25.02** (+2.10)	**50.54** (+0.77)
CECL (ViViT)	18.05 (+1.08)	10.06 (-0.27)	37.31 (+1.62)	57.26 (+1.15)	12.92 (-1.28)	11.39 (+0.09)	20.43 (-2.41)	47.72 (-0.41)
CECL (VideoMAE)	14.29 (+1.36)	11.18 (-0.29)	22.08 (+1.08)	47.26 (+0.01)	13.04 (+0.47)	11.38 (-0.12)	20.18 (+2.08)	47.09 (+1.45)
CECL (SigLIP)	17.94 (+0.27)	10.00 (+0.43)	35.88 (-3.90)	57.11 (-2.29)	13.47 (+2.45)	11.14 (-0.23)	24.49 (+4.59)	49.64 (+2.72)

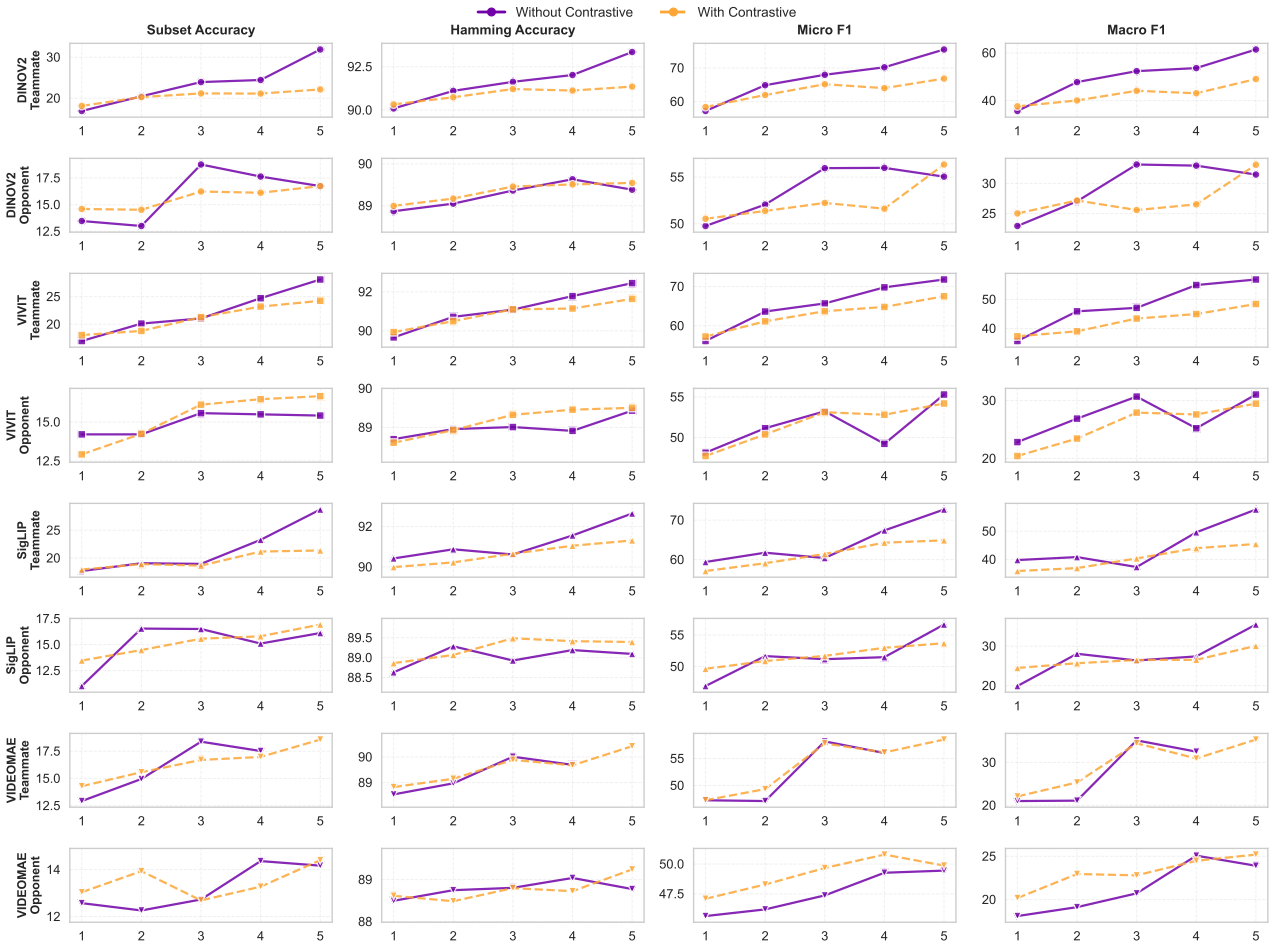


Figure 4: Performance comparison of CECL vs. baseline across different visual encoders (DINOv2, ViViT, VideoMAE, SigLIP) as the number of agent POVs varies from 1 to full team size of 5. CECL shows substantial gains at lower POV counts (1-2 agents) but diminishing returns as team coverage increases.

Entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}[y_{ij}\log(\hat{p}_{ij}) + (1-y_{ij})\log(1-\hat{p}_{ij})]$$

where N is the batch size, $L = 23$ is the number of location classes, $y_{ij} \in \{0, 1\}$ are ground truth labels, and $\hat{p}_{ij} \in [0, 1]$ are predicted probabilities. The total loss combining with CECL (eq. 1) is:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{CECL}} + \mathcal{L}_{\text{BCE}}$$

where λ is a balancing weighting coefficient.

5.2 Quantitative Evaluation

Single POV Evaluation. Table 2 shows the performance comparison of CECL vs. off-the-shelf state-of-the-art visual models on Teammate Location Nowcast and Enemy Location Nowcast tasks across 4 multi-label classification metrics, using only single-agent POV visual embeddings as input. The results demonstrate that the CECL learning objective generally boosts performance at the single-agent POV level across most model and metric combinations, with a mean absolute gain of 1.45 percentage points over all baselines.

Multi-POV Evaluation. We further evaluate the effectiveness of CECL by varying the number of agent POVs available during inference. Figure 4 shows performance comparison between models with and without CECL training across different visual encoders (DINOv2, ViViT, VideoMAE, SigLIP) as the number of selected agents increases from 1 POV to the full team size of 5 POVs.

The results demonstrate a clear pattern: CECL provides substantial performance gains when only 1-2 agent POVs are available. However, as the number of POVs increases toward full team observation (5 agents), the advantage of CECL diminishes. This is expected because when full team observations are available, concatenation-based aggregation already preserves complete information from all viewpoints. In contrast, in low-POV settings, observations are inherently partial and limited, making cross-egocentric alignment more valuable.

CECL's slight underperformance at full-team settings compared to non-contrastive baselines can be attributed to the contrastive objective. While the contrastive objective enhances information sharing among POVs, the training process may lead to a more concentrated embedding space, potentially causing some information loss from the original visual embedding space. This is also indicated by the t-SNE visualization in Figure 5, where the post-contrastive embeddings form a thinner, more compact manifold. Importantly, CECL is specifically designed for low-POV scenarios—settings in which only a single player's first-person view is available during inference, as would be the case in real gameplay. This could enabling single agent to infer team-level context from its own perspective and paving the way for team-aware AI teammates that reason about collective dynamics from individual experience.

5.3 Ablation Study

Because the sigmoid loss is sensitive to the large imbalance of positive and negative pairs in a batch, we ablate the values of the counter parameters b and t in the sigmoid loss and validate our design choice of the parameters as mentioned in Section 4.3. As shown in Table 3, without proper bias initialization ($b = 0$), the model struggles to learn effective representations, while a bias of -3 and temperature of log 10 generally leads to the best performance.

Table 3: Impact of bias term and temperature on contrastive learning performance on DINOv2. The bias b addresses the imbalance between positive and negative pairs in the batch.

b	t	Teammate		Enemy	
		Sub.Acc	Ham.Dist	Sub.Acc	Ham.Dist
n/a	log 10	14.13	11.37	11.44	11.64
0	log 10	13.90	11.42	9.07	11.70
0	log 1	14.68	10.51	12.15	11.36
-3	log 10	**17.01**	**10.38**	**12.38**	11.29
-3	log 1	13.47	11.17	11.68	10.98
-10	log 10	15.77	10.49	10.67	**10.96**
-10	log 1	1.40	12.86	1.40	12.92

5.4 Qualitative Evaluation

We visualize the effectiveness of our cross-egocentric contrastive learning approach using t-SNE dimensionality reduction [34]. Figure 5 compares the embedding space before (bottom row) and after (top row) contrastive training, with columns showing different coloring schemes: location (left), time (middle), and team (right). The analysis reveals substantial improvements in representation quality after CECL training, with location-based clustering showing an 8.87% increase in separation ratio (from 1.272 to 1.385) and team-based clustering demonstrating a 15.86% improvement (from 1.040 to 1.205). The temporal visualization exhibits clearer cluster margins and smoother gradients after training, indicating that CECL preserves meaningful temporal progression while enhancing spatial clustering.

6 DISCUSSION

Our findings suggest that cross-egocentric representation learning offers a promising new paradigm for understanding multi-agent coordination in visually rich, partially observable domains. By aligning teammates' egocentric video streams, CECL enables agents to internalize shared latent representations that approximate a global team state without requiring explicit communication or centralized supervision. This implicit coordination mechanism parallels how human teammates achieve joint situational awareness through mutual observation, anticipation, and inference.

A notable observation is that CECL's benefits are most pronounced under low observability, when only one or two egocentric perspectives are available. This suggests that contrastive alignment can act as an implicit communication channel: by enforcing representational consistency across teammates, the model learns to reconstruct unobserved aspects of the shared environment. In essence, CECL allows an agent to "imagine" what its teammates might be perceiving, resembling theory-of-mind reasoning. However, this implicit communication introduces a trade-off as representation compression that can slightly reduce expressivity when

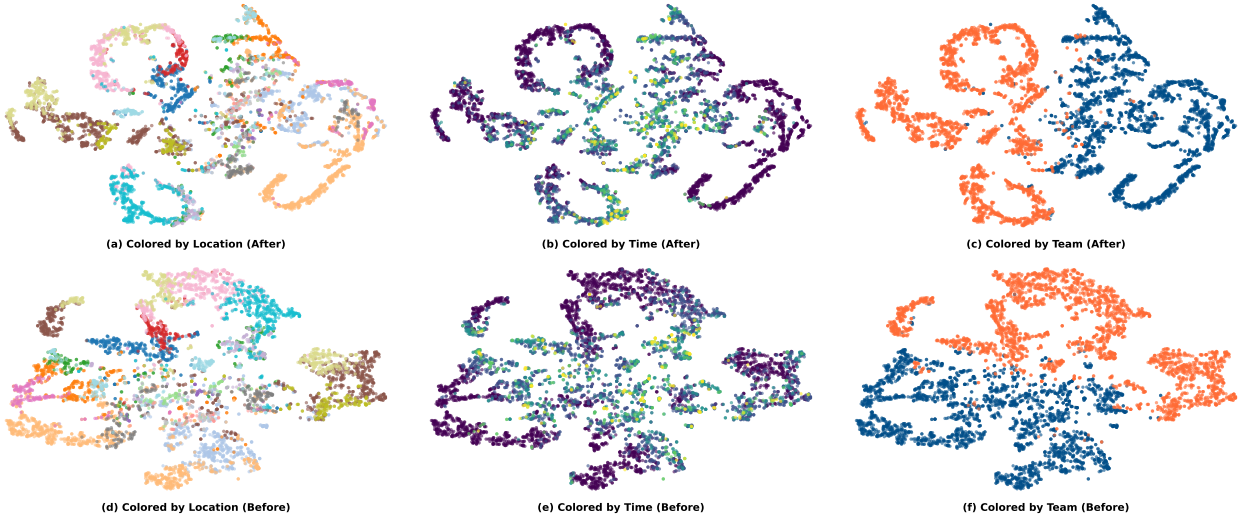


Figure 5: t-SNE visualization of learned embeddings before (bottom row) and after (top row) cross-egocentric contrastive learning. Each point represents a video clip embedding. Columns show different coloring schemes: location (left), time (middle), and team (right). After contrastive training, embeddings demonstrate substantially improved clustering with 8.87% increase in location-based separation and 15.86% increase in team-based separation, while maintaining smooth temporal gradients.

full team visibility is available. Future research could explore hybrid contrastive-reconstruction objectives to balance shared and individual encoding fidelity.

Interestingly, CECL did not consistently improve performance on the recent V-JEPA2 0.3B model [2] in our experimentations, in contrast to other models. One possible explanation is that V-JEPA2's high model capacity and predictive pretraining already encode strong spatiotemporal coherence, leaving limited headroom for additional contrastive regularization. Alternatively, such high capacity models (with over 3 times more parameters than the rest of the models) may require longer training horizons, refined temperature scaling, or customized optimization schedules.

In addition, our current framework aligns teammates' egocentric representations within the same team and timestep, but does not jointly align all teammates and opponents simultaneously. While this design isolates cooperative information sharing, jointly modeling both sides of the encounter could provide richer tactical structure by capturing inter-team dependencies and adversarial intent. Extending CECL to a fully cross-team alignment setting, where all agents' viewpoints at the same timestep are contrasted within a unified embedding space, represents a promising future direction.

Although our study centers on competitive gaming, the underlying principles extend naturally to other domains involving human-AI teaming under partial observability, such as collaborative robotics, defense operations, and autonomous vehicle fleets. CECL's ability to infer unobserved teammate states from individual sensory inputs points toward scalable models of shared situational awareness in mixed human-machine systems. In these contexts, cross-egocentric alignment could serve as a foundation for interpretable, coordination-aware world models that integrate perception, prediction, and planning.

We identify several promising directions for future work. First, while our experiments focused on contrastive self-supervision for learning team-level awareness, alternative paradigms such as masked video modeling could complement this approach. Inspired by masked language modeling in NLP, future extensions might randomly mask an agent's viewpoint and train a video generator to reconstruct missing perspectives, enhancing cross-agent imagination. Second, cross-egocentric data exist abundantly in real-world human teamwork scenarios—for example, from body-worn or vehicle-mounted cameras—offering opportunities to model multi-human coordination in naturalistic settings. Third, integrating CECL-trained representations into controllable actor agents remains an open direction, enabling direct evaluation of CECL's influence on embodied decision-making and AI-human collaboration.

7 CONCLUSION

We introduced **X-Ego-CS**, a benchmark dataset for cross-egocentric multi-agent video understanding in professional esports, and proposed **Cross-Ego Contrastive Learning (CECL)**, a self-supervised framework that aligns teammates' first-person visual streams to acquire team-level situational awareness. Through extensive experiments, we demonstrated that CECL enhances agents' ability to infer both teammate and opponent positions from limited egocentric views, particularly under partial observability, highlighting the potential of cross-egocentric alignment as a scalable mechanism for fostering implicit coordination and shared tactical understanding without explicit communication or supervision. More broadly, this work establishes a foundation for collective perception and reasoning in multi-agent systems, bridging self-supervised representation learning and human-AI teaming, and opens new pathways for understanding and reproducing team dynamics in complex, real-time environments through cross-perspective alignment.

ACKNOWLEDGMENTS

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. 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.

REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6836–6846.

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985* (2025).

[3] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2019. Emergent tool use from multi-agent autocurricula. In *International conference on learning representations*.

[4] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 33.

[5] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[6] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.

[7] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. In *Advances in neural information processing systems*, Vol. 32.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[9] Tom Decroos, Jan Van Haaren, and Jesse Davis. 2018. Automatic discovery of tactics in spatio-temporal soccer match data. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 223–232.

[10] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. 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.

[11] David Durst, Feng Xie, Vishnu Sarukkai, Brennan Shacklett, Iuri Frosio, Chen Tessler, Joohwan Kim, Carly Taylor, Gilbert Bernstein, Sanjiban Choudhury, et al. 2024. Learning to Move Like Professional Counter-Strike Players. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15173.

[12] FACEIT. 2025. FACEIT Platform. https://www.faceit.com/en. Accessed: 2025-08-04.

[13] Marios Fanourakis and Guillaume Chanel. 2025. AMuCS: Affective multimodal Counter-Strike video game dataset. *Scientific Data* 12, 1 (2025), 1325.

[14] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1711–1721.

[15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.

[16] Edward Gu, Ho Chit Siu, Melanie Platt, Isabelle Hurley, Jaime Peña, and Rohan Paleja. 2025. Enabling Rapid Shared Human-AI Mental Model Alignment via the After-Action Review. *arXiv preprint arXiv:2503.19607* (2025).

[17] Xiaofan Gu, Xinwei Xue, and Feng Wang. 2020. Fine-grained action recognition on a novel basketball dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2563–2567.

[18] Nikolos Gurney, Stacy Marsella, Volkan Ustun, and David V Pynadath. 2021. Operationalizing theories of theory of mind: A survey. In *AAAI Fall Symposium*. Springer, 3–20.

[19] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.

[20] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104* (2023).

[21] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. 2025. Training Agents Inside of Scalable World Models. *arXiv preprint arXiv:2509.24527* (2025).

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[23] Zican Hu, Zongzhang Zhang, Huaxiong Li, Chunlin Chen, Hongyu Ding, and Zhi Wang. 2023. Attention-guided contrastive role representations for multi-agent reinforcement learning. *arXiv preprint arXiv:2312.04819* (2023).

[24] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. 2016. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences* 20, 8 (2016), 589–604.

[25] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 4501–4510.

[26] LaihoE. 2025. demoparser: Counter-Strike 2 replay parser for Python and JavaScript. https://github.com/LaihoE/demoparser. Accessed: 2025-09-14.

[27] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. 2006. A tutorial on energy-based learning. *Predicting structured data* 1, 0 (2006).

[28] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037* (2017).

[29] Tianxu Li, Kun Zhu, Juan Li, and Yang Zhang. 2024. Learning distinguishable trajectory representation with contrastive loss. *Advances in Neural Information Processing Systems* 37 (2024), 64454–64478.

[30] Yat Long Lo, Biswa Sengupta, Jakob Foerster, and Michael Noukhovitch. 2023. Learning multi-agent communication with contrastive learning. *arXiv preprint arXiv:2307.01403* (2023).

[31] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[32] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).

[33] Wenhao Ma, Yu-Chen Chang, Jie Yang, Yu-Kai Wang, and Chin-Teng Lin. 2025. Contrastive learning-based agent modeling for deep reinforcement learning. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2025).

[34] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[35] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[36] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5074–5085.

[37] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[38] NVIDIA Corporation. 2025. NVIDIA APP. https://www.nvidia.com/en-us/geforce/geforce-experience/shadowplay/. Accessed: 2025-09-14.

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[41] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*. PMLR, 4218–4227.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[43] Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Towards universal soccer video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8384–8394.

[44] Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. 2024. Matchtime: Towards automatic soccer game commentary generation. *arXiv preprint arXiv:2406.18530* (2024).

[45] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research* 21, 178 (2020), 1–51.

[46] Sugandha Sharma, Guy Davidson, Khimya Khetarpal, Anssi Kanervisto, Udit Arora, Katja Hofmann, and Ida Momennejad. 2024. Toward human-ai alignment in large-scale multi-player games. *arXiv preprint arXiv:2402.03575* (2024).

[47] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[48] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[49] Ho Chit Siu, Jaime Peña, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chang, and Ross Allen. 2021. Evaluation of human-ai teams for learned and rule-based agents in hanabi. *Advances in Neural Information Processing Systems* 34 (2021), 16183–16195.

[50] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2446–2454.

[51] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

[52] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *nature* 575, 7782 (2019), 350–354.

[53] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).

[54] Yunzhe Wang, Volkan Ustun, and Chris McGroarty. 2025. A data-driven discretized CS:GO simulation environment to facilitate strategic multi-agent planning research. In *Proceedings of the 2025 Winter Simulation Conference (WSC)*. IEEE, Los Angeles, CA, USA.

[55] Zhe Wang, Petar Veličković, Daniel Hennes, Nenad Tomašev, Laurel Prince, Michael Kaisers, Yoram Bachrach, Romuald Elie, Li Kevin Wenliang, Federico Piccinini, et al. 2024. TacticAI: an AI assistant for football tactics. *Nature communications* 15, 1 (2024), 1906.

[56] Peter Xenopoulos and Claudio Silva. 2022. Esta: An esports trajectory and action dataset. *arXiv preprint arXiv:2209.09861* (2022).

[57] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean field multi-agent reinforcement learning. In *International conference on machine learning*. PMLR, 5571–5580.

[58] Peihong Yu, Manav Mishra, Syed Zaidi, and Pratap Tokekar. 2025. TACTIC: Task-Agnostic Contrastive pre-Training for Inter-Agent Communication. *arXiv preprint arXiv:2501.02174* (2025).

[59] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

[60] Lingyu Zhang, Zhengran Ji, and Boyuan Chen. 2024. Crew: Facilitating human-ai teaming research. *arXiv preprint arXiv:2408.00170* (2024).

[61] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[62] Shao Zhang, Xihuai Wang, Wenhao Zhang, Chaoran Li, Junru Song, Tingyu Li, Lin Qiu, Xuezhi Cao, Xunliang Cai, Wen Yao, et al. 2025. Leveraging dual process theory in language agent framework for real-time simultaneous human-AI collaboration. *arXiv preprint arXiv:2502.11882* (2025).