Temporal Action Representation Learning for Aerial Maneuvering and Resource-Aware Decision-Making

Hoseong Jung, Sungil Son, Daesol Cho, Jonghae Park, Changhyun Choi, H. Jin Kim*
Lab for Autonomous Robotics Research, Seoul National University

***Abstract*—A fully autonomous agent should reason about how to deploy limited resources effectively in dynamic and uncertain environments. Despite the focus on learning to act under such constraints, the tactical use of resources in fast-evolving scenarios (e.g., air combat) remains underexplored. Addressing this challenge requires modeling how resource usage unfolds over time and influences future behavior. To this end, we explore self-supervised learning approaches tailored to modeling the causal dependency and temporal relationship between these interrelated processes. Specifically, we introduce TART, a Temporal Action contrastive learning approach that facilitates semantic alignment between Resource control and Tactical maneuvers. TART learns via contrastive learning based on a mutual information objective, carefully designed to account for both forward and backward dependencies embedded within such dynamics. These learned representations are quantized into discrete codebook entries that serve as inputs to the policy, enabling us to model the multiple tactical modes in downstream tasks. To empirically assess our method, we present an air combat simulation environment where tactical resource allocation is essential for mission success, using customized scenarios of varying difficulty to compare against baseline approaches. Extensive experiments demonstrate the effectiveness of TART in the use of limited resources and superior performance in generating tactical maneuvers.**

I. INTRODUCTION

General autonomous robotic systems are required to operate under limited resources due to constraints such as communication bandwidth, fuel, or cost [1, 2]. In static environments, this challenge lies in the resource allocation problem involving the assignment of resources or tasks to agents and high-level operational planning, here referred to as *strategic allocation*. Within such settings, optimization methods guided by appropriate constraints and planning with accurate system models have been shown to be effective. In contrast, in dynamic environments (e.g., air combat) where rapid decisions are necessary, effective resource usage depends on detailed tactical decisions, which we refer to as *tactical allocation*. Here, agents must reason not only about how to use their resources but also how such decisions influence subsequent maneuvers.

In such settings, hybrid action spaces naturally arise: discrete actions govern resource usage (e.g., firing missiles), while continuous actions control low-level behaviors [3, 5]. While reinforcement learning (RL) is well-suited to these problems due to its adaptability, effective tactical resource allocation presents three major challenges: **causal dependency**, **temporal understanding**, and **multi-modality** as shown in Fig. 1. Specifically, agents should understand the causal effects of resource usage, reason over future results of current actions, and account for the multi-modal nature of tactical behavior.

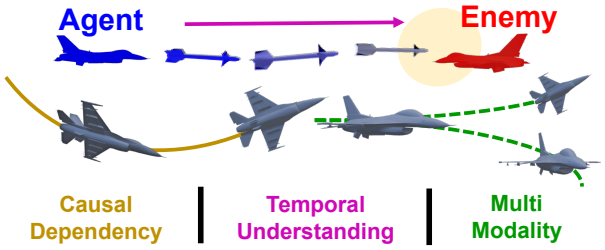

Fig. 1: Key challenges in tactical resource allocation in air combat. The agent fires a missile and must select an appropriate subsequent maneuver, showing the **causal dependency** between resource usage and following actions. Implicit reasoning about the missile's effectiveness reflects the need for **temporal understanding**. Multiple valid follow-up maneuvers highlight the **multi-modal** nature of tactical decision-making.

For example, in air combat, deploying a high-impact resource like a missile may lead to a variety of valid follow-up maneuvers (see Fig. 3) depending on the evolving situation. The agent must anticipate possible outcomes and choose the most appropriate tactical maneuvers to fully exploit the advantage.

To address these challenges, we hypothesize that grouping semantically similar actions in a latent space facilitates a better understanding of temporal relationships critical for tactical decision-making. In this paper, we introduce **T**emporal **A**ction representation learning for **R**esource control and **T**actical maneuver generation (TART). TART builds upon a mutual information objective designed to reflect both forward and backward dependencies between resource utilization and subsequent maneuvers. To better capture fine-grained temporal relationships, we extend this objective with a contrastive loss that enforces similarity constraints across trajectories. The resulting representations are quantized into discrete codes via vector quantization, which are then used to condition the policy to produce context-aware hybrid actions.

We construct a high-fidelity air-to-air combat environment to evaluate tactical decision-making under discrete-continuous action settings. The environment simulates engagements between fighter jets with realistic weapon and countermeasure systems. The agent controls an F-16 aircraft equipped with missiles, gunfire, and defensive systems such as chaff and flares. To isolate the tactical aspects of decision-making, the initial scenario configuration (e.g., position, resource availability) is fixed across episodes, allowing the agent to focus exclusively on resource control and maneuver generation. Through experiments, we demonstrate that TART outperforms baselines, and analysis of the representations shows that TART captures diverse tactical modes.

II. BACKGROUND

A. Parameterized Action Markov Decision Process

In this paper, we build on a Parameterized Action Markov Decision Process (PAMDP) $< \mathcal{S}, \mathcal{H}, \mathcal{P}, \gamma, \mathcal{R} >$, defined with a state space $\mathcal{S}$, parameterized action space $\mathcal{H}$, transition function $\mathcal{P} : \mathcal{S} \times \mathcal{H} \times \mathcal{S} \to \mathbb{R}$, discounted factor $\gamma \in [0, 1)$, and reward function $\mathcal{R} : \mathcal{S} \times \mathcal{H} \to \mathbb{R}$ [3]. Specifically, we extend the standard PAMDP framework to incorporate a discrete-continuous hybrid action space $\mathcal{H}$ defined as:

$$\mathcal{H} = \{(k, x_k) \,|\, k = (a_d; a_c), a_d \in \mathcal{A}_d, a_c \in \mathcal{A}_c, x_k \in \mathcal{X}_k\} \quad (1)$$

where $\mathcal{A}_d = \{a_{d,1}, ..., a_{d,m}\}$ is the m-dimension discrete action set, $\mathcal{A}_c = \{a_{c,1}, ..., a_{c,n}\}$ is the n-dimension continuous action set, and $\mathcal{X}_k$ is the corresponding parameter set for each hybrid action identifier k. Note that discrete action controls m types of resources, while continuous action controls n types of low-level actions.

Such hybrid action spaces commonly emerge in real-world tasks [3, 4, 5], which attract attention to the RL community. The complexity of hybrid action spaces and the interdependence among their components necessitate RL agents to model these dependencies while preserving scalability and stationarity [5]. To address these challenges, recent works propose Q-learning [3], actor-critic [4] approaches to overcome these challenges. Notably, Li *et al*. proposed HyAR [5] that construct a latent embedding space for single-step actions via conditional Variational Autoencoder (cVAE), successfully resolving the redundancy issue in the enlarged action space. Nevertheless, incorporating a separate generative model causes considerable complexity and error accumulations, and single-step representation remains insufficient to capture temporal relationships.

III. METHODS

This section mainly introduces the overall theoretical framework and the details of the implementation of TART. The key idea of TART is to learn diverse tactical representations that are informative about the hybrid action space. To enable TART to effectively use limited resources, we introduce bidirectional objectives based on mutual information, along with a practical contrastive learning loss function. The disentangled representations are transformed into discrete, interpretable codebooks, which condition a policy network to generate a multi-modal hybrid action distribution. Fig. 2 provides an overview of the proposed method.

A. Bidirectional Objective for Representation Learning

We begin by presenting the temporal contrastive learning objectives of TART. The guiding principle of our method is to learn state and action representations that capture temporal relationships and dependencies essential for learning the optimal policy. To quantify the degree of the relationship, we employ mutual information, denoted $\mathcal{I}(X; Y)$, which is a reparameterization-invariant measure of dependency:

$$\mathcal{I}(X; Y) = \mathbb{E}_{p(x,y)} \log \frac{p(x, y)}{p(x)p(y)} = H(X) - H(X|Y), \quad (2)$$

where X and Y represent either raw samples or stochastic representations from a data distribution. We define the state representations as $z_t = \phi(s_t)$, and the hybrid action representation as $u_{t,d} = \psi(a_{t,d})$ for discrete actions and $u_{t,c} = \psi(a_{t,c})$ for continuous actions, where s_t, $a_{t,d}$, and $a_{t,c}$ denote the raw state and hybrid action components. Here, ϕ and ψ serve as encoders for the state s_t, and the action components $a_{t,d}$ and $a_{t,c}$. We then propose two MI-based objectives, each distinguished by their directional dependence and formally defined in Equations (3) and (4). K denotes a fixed hyperparameter for the predicted horizon.

Forward dynamics

$$\mathbb{J}_{fwd} = \mathcal{I}(u_{t+k,c}; [z_t, u_{t:t+K-1}]), \quad (3)$$

Inverse dynamics

$$\mathbb{J}_{inv} = \mathcal{I}(u_{t,d}; [z_{t+K}, u_{t:t+K-1} \mid z_t]). \quad (4)$$

Given an objective $\mathbb{J}$, we define the optimal set of state and action encoders $\Phi_{\mathbb{J}}$ as those that jointly maximize $\mathbb{J}$:

$$\Phi_{\mathbb{J}} = \{(\phi, \psi) \in \mathcal{F} \times \mathcal{G} \mid (\phi, \psi) \in \arg\max_{\phi', \psi'} \mathbb{J}(\phi', \psi')\}, \quad (5)$$

where $\mathcal{F}$ and $\mathcal{G}$ denote the function classes from which the state encoder ϕ and the action encoder ψ are chosen, respectively. By extending the theorem from Rakely *et al.* [6], we guarantee that if $(\phi, \psi) \in \Phi_{\mathbb{J}}$, then the corresponding state-action pair (s_t, a_t) leads to accurate estimate of the optimal action-value function $Q^*(s_t, a_t)$.

B. Contrastive Learning of Latent Action Representations

Since mutual information is typically intractable to compute directly, we instead adopt InfoNCE [7] loss to derive and optimize a tractable lower bound:

$$\mathcal{I}(X; Y) \geq \log(N) - \mathcal{L}_C, \quad (6)$$

where N denotes the number of samples and $\mathcal{L}_C$ is the InfoNCE loss. We then organize the state-action representations into an anchor w, a positive sample w^+, and negative samples w^-. The InfoNCE loss encourages the anchor to be similar to the positive while dissimilar to the negatives:

$$\mathcal{L}_C = -\log \frac{\exp(\text{sim}(w, w^+))}{\exp(\text{sim}(w, w^+)) + \sum_{w^-} \exp(\text{sim}(w, w^-))}, \quad (7)$$

where $\text{sim}(w_i, w_j) = w_i^\top W w_j$ is a learnable similarity function with parameter W.

Given two sets of state-action representations at time step t, we define the contrastive loss separately for the forward and inverse dynamics objectives. For the forward dynamics objective, we set the anchor as $w = [z_t, u_{t:t+K-1,d}]$, the positive as $w^+ = u_{t+k,c}$, and the negative w^- as continuous actions sampled from other trajectories or time steps. For the inverse dynamics objective, the anchor is $w = [z_{t+K}, u_{t:t+K-1,c}]$ conditioned on z_t; the positive is $w^+ = u_{t,d}$, and the negatives w^- are discrete actions sampled from the batch except $u_{t,d}$. Both objectives capture tactical modes, encouraging representation clustering via contrastive learning. Following TACO [8], we augment positive pairs using temporally adjacent segments within the same episode to improve sample efficiency.

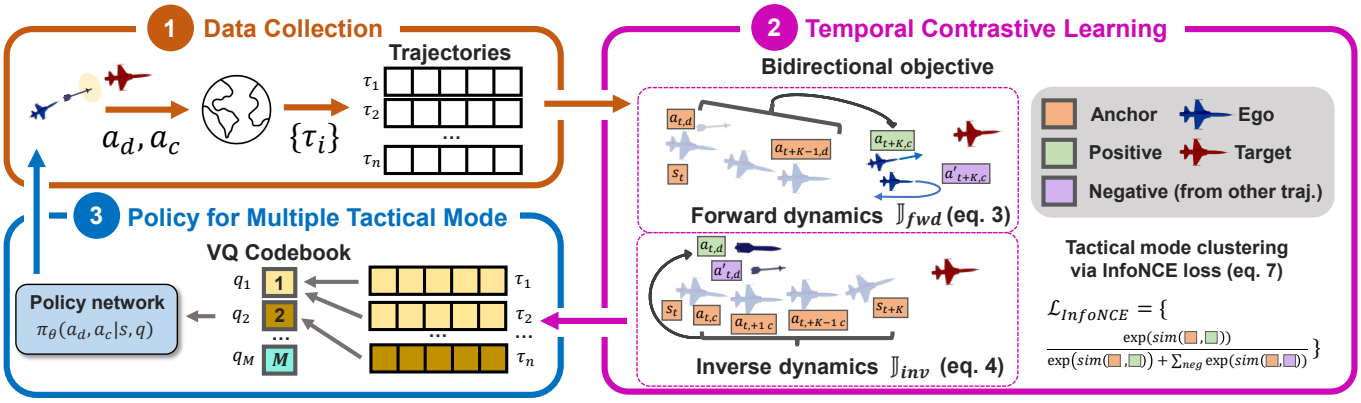

Fig. 2: Overview of TART: (1) The agent interacts with the environment and collects a set of trajectories $\{\tau_i\}$. (2) A bidirectional objective guides the clustering of given trajectories into multiple tactical modes through contrastive learning. The resulting distinct modes are then mapped to discrete vectors via vector quantization (VQ). (3) A policy network distinguishes between the modes and generates multi-modal hybrid action distributions accordingly.

C. Learning Discrete Tactical Modes via Vector Quantization

To exploit the clustered action representations obtained via contrastive learning, we aim to represent them as discrete tactical modes that guide the policy in generating diverse action distributions [9]. Let $\mathcal{C} = \{c_1, ..., c_M\}$ denote the set of learnable codebook entries, where each $c_i \in \mathbb{R}^d$. Given a state-action trajectory $\tau_t = \{(s_{t'}, a_{t'}), ..., (s_t, a_t)\}$, where $t' = t - K + 1$, we obtain state and action embedding $\{z_{t'}, ..., z_t\}$ and $\{u_{t'}, ..., u_t\}$ via encoders ϕ and ψ as shown in Section III.A. These sequences are fed into a learnable function f to produce a trajectory-level representation $\kappa = f(\tau_t)$. The embedding κ is quantized by assigning it to the nearest codebook entry c_i based on Euclidean distance, resulting in a discrete latent code q. This quantized code serves as a tactical mode and is used to condition the policy $\pi_\theta(a_c, a_d|s, q)$.

To learn effective quantized representations, we employ the standard vector quantization objective, consisting of a reconstruction loss and a commitment loss:

$$\mathcal{L}_{VQ} = ||\kappa - \hat{\kappa}||^2 + ||\text{sg}[\kappa] - c_i||^2 + \beta||\kappa - \text{sg}[c_i]||^2, \quad (8)$$

where $\hat{\kappa}$ is the reconstruction of the embedding from the codebook, $\text{sg}[\cdot]$ denotes the stop-gradient operator, and β is a hyperparameter that balances the commitment strength. The VQ encoder is optimized to produce embeddings κ that closely match their assigned codebook entries, while the codebook itself adapts to reflect the encoder outputs. At inference time, a trajectory segment is processed by VQ encoder and quantized into its corresponding code q, which conditions the policy to produce context-aware multi-modal hybrid actions.

IV. EXPERIMENTS

This section presents the empirical evaluation of TART in a custom-designed air combat environment, conducted in an online RL setting. Throughout all experiments, we use Proximal Policy Optimization (PPO) [10] for the backbone algorithm and fix the prediction horizon parameter to $K = 10$ and the commitment loss coefficient to $\beta = 0.25$. It is worth noting that TART is designed to be compatible with any online RL algorithm that supports hybrid action spaces.

A. Experimental Setup in Aerial Tactical Environments

We build a custom air combat environment upon the Light Aircraft Game (LAG) [11] and NeuralPlane [12], simulating F-16 fighter jets equipped with weapons and countermeasures. Offensive systems include AIM-9M and AIM-120B missiles with pursuit-point guidance and a gun system modeled after the AlphaDogfight Trial [13]. Defensive systems include chaff and flares to evade incoming missiles. The aircraft is controlled via a hierarchical architecture following [14], where our online RL agent acts as the high-level combat policy.

The environment follows the standard PAMDP framework. Following prior work [15], the state space $\mathcal{S}$ includes aircraft dynamics and geometric configurations, remaining weapons, and countermeasures. Discrete actions $\mathcal{A}_d$ include five options: AIM-9M, AIM-120B, gunfire, chaff and flare, while continuous actions $\mathcal{A}_c$ control three aircraft maneuver parameters. The transition function $\mathcal{P}$ is defined by the simulator's physics [16], and the reward function $\mathcal{R}$ is designed to be sparse: agents receive positive rewards only when shooting down the opponent and get negative rewards for being shot down, crashing, or wasting munitions. To evaluate performance, we define two hand-scripted opponents based on manually specified missile launch conditions and maneuvering logic: *Easy, Medium*. Agents are initialized with randomized altitudes, positions, and orientations at the start of each episode and assessed by their success in defeating these opponents.

B. Experimental Results

We compare TART against three reinforcement learning baselines for hybrid action spaces: IPPO [4], IPPO-prior, and HyAR [5]. IPPO extends PPO by separately handling discrete and continuous actions. IPPO-prior augments domain knowledge via a Bernoulli prior over discrete actions to improve missile launch decisions. HyAR employs a conditional Variational AutoEncoder (cVAE), enabling the generation of continuous actions conditioned on discrete actions. As shown in Table 1, TART achieves consistently higher win rates against both hand-scripted opponents, demonstrating superior tactical decision-making in hybrid action settings.

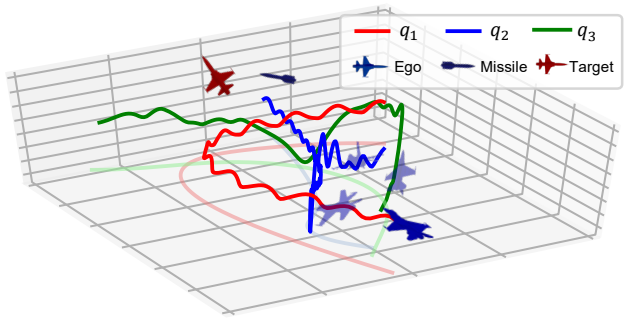

Fig. 3: Visualization of distinct tactical modes learned by TART. Agents begin from similar spatial configurations but differ in conditioning trajectories (e.g., past actions, remaining missiles). As a result, they are assigned to different VQ codebook entries, inducing maneuvers with varying aspect angles. Shaded regions show 2D ground-plane projections.

TABLE I: Win rates against hand-scripted opponents in an air combat environment, evaluated over 3 random seeds.

Method	Easy	Medium
IPPO	74.0 $\pm$ 8.04	58.3 $\pm$ 11.14
IPPO-prior	88.3 $\pm$ 3.39	68.6 $\pm$ 7.58
HyAR	66.0 $\pm$ 6.37	69.3 $\pm$ 4.49
TART	**97.6** $\pm$ **1.69**	**78.6** $\pm$ **2.4**

To evaluate whether vector-quantized representation captures diverse tactical maneuvers, we analyze the *Medium* task. Agents start from similar spatial configurations but differ in two aspects: (1) the number of remaining missiles, which serves as conditioning state input, and (2) the trajectory history leading up to the current state. We examine how these factors lead to the selection of VQ codebook indices and produce distinct trajectories. As shown in Fig. 3, the resulting trajectories reflect semantically meaningful tactical variations. Notably, selection of q_1 correlates with lower missile availability, while q_2 corresponds to states with more remaining missiles.

V. CONCLUSION

In this paper, we introduce TART, a plug-and-play module for hybrid action space reinforcement learning that enables effective resource control and tactical maneuver generation. TART is compatible with a variety of online RL algorithms and performs well in a custom air combat environment. For future work, we plan to develop a more challenging hand-scripted opponent (*Difficult*) and incorporate self-play strategies to further improve the agent's tactical capabilities. We also aim to explore alternative methods for multi-modal action generation to better capture the diversity of tactical behaviors. These directions offer promising avenues for applying TART to broader resource-constrained decision-making tasks.

REFERENCES

[1] B. P. Gerkey and M. J. Matarić, "A formal analysis and taxonomy of task allocation in multi-robot systems," *The International Journal of Robotics Research*, vol. 23, no. 9, pp. 939–954, 2004.

[2] M. Afrin, J. Jin, A. Rahman, *et al.*, "Resource allocation and service provisioning in multi-agent cloud robotics: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 842–870, 2021.

[3] W. Masson, P. Ranchod, and G. Konidaris, "Reinforcement learning with parameterized actions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[4] Z. Fan, R. Su, W. Zhang, and Y. Yu, "Hybrid actor-critic reinforcement learning in parameterized action space," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019, pp. 2279–2285.

[5] B. Li, H. Tang, Y. Zheng, *et al.*, "Hyar: Addressing discrete-continuous action reinforcement learning via action representation," in *Proceedings of the International Conference on Learning Representations*, 2022.

[6] K. Rakelly, A. Gupta, C. Florensa, and S. Levine, "Which mutual-information representation learning objectives are sufficient for control?," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26345–26357, 2021.

[7] A. Van Den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[8] R. Zheng, X. Wang, Y. Sun, *et al.*, "TACO: Temporal latent action-driven contrastive loss for visual reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 48203–48225, 2023.

[9] A. Van Den Oord and O. Vinyals, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] J. Schulman, F. Wolski, P. Dhariwal, *et al.*, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[11] Q. Liu, Y. Jiang, and X. Ma, "Light aircraft game: A lightweight, scalable, gym-wrapped aircraft competitive environment with baseline reinforcement learning algorithms," 2022, URL https://github.com/liuqh16/CloseAirCombat

[12] C. Xue, Q. Liu, *et al.*, "NeuralPlane: An efficiently parallelizable platform for fixed-wing aircraft control with reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 96939–96962.

[13] A. P. Pope, J. S. Ide, D. Mićović, *et al.*, "Hierarchical reinforcement learning for air combat at DARPA's AlphaDogfight Trials," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 6, pp. 1371–1385, 2022.

[14] J. Chai, W. Chen, Y. Zhu, *et al.*, "A hierarchical deep reinforcement learning for 6-DOF UCAV air-to-air combat," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 9, pp. 5417–5429, 2023.

[15] J. Bae, H. Jung, S. Kim, *et al.*, "Deep reinforcement learning-based air-to-air combat maneuver generation in a realistic environment," *IEEE Access*, vol. 11, no. 1, pp. 26427–26440, 2023.

[16] J. Berndt, "JSBSim: An open source flight dynamics model in C++," in *Proceedings of AIAA Modeling and Simulation Technologies Conference and Exhibit*, 2004, pp. 4923.