Digital Intelligence PRESS

Journal of Digital Frontier

https://doi.org/10.67541/jdf2612

Research Article **Volume 1, Issue 2 (Sep 2026)** *Open Access*

Research on Simulation and Path Optimization of University Emergency Evacuation Based on Multi-Agent Reinforcement Learning

Linna Yang *

School of Economics and Management, Lanzhou Institute of Technology, Lanzhou, Gansu, China

***Corresponding author:** Linna Yang. Email: yangln@lzit.edu.cn

Abstract

To address the spatial misalignment between local observations and global bottleneck consequences in university emergency evacuation, as well as the difficulty of sparse team rewards in guiding diversion behaviors, the Topological Awareness Hierarchical State Encoder (TAHSE) and Bottleneck Sensitive Reward Decomposition Mechanism (BSRD) are proposed. These two mechanisms are coupled into a unified architecture within a centralized training-decentralized execution framework. TAHSE aggregates the features of the local neighborhood graph using attention and uses learnable bottleneck tokens as the keys and values for the second layer's cross-domain attention, enabling agents to obtain the compressed prior of the global bottleneck solely based on their own observations and pre-stored bottleneck graphs, reducing the computational complexity from the square level of the full graph attention to the linear level. BSRD decomposes a single negative time team reward into three components: progress, bottleneck pressure, and coordination completion. It triggers pressure penalties with a warning threshold lower than the safety upper limit and uses an adaptive coefficient to weigh the safety constraints as the difficulty of the scenario increases. Twenty seed repetition experiments were conducted in three-tier campus scenarios (500/1500/3000 agents) and structured perturbation protocols: The gap ratio of the proposed method relative to the upper bound reference in the three scenarios was controlled within 15%, while the MAPPO baseline increased from 18.4% to 31.5% as the scale expanded; the peak density effect under topological perturbation reached 1.42 ($p<0.001$), and the triggering threshold for diversion shifted systematically towards the bottleneck nodes with reduced capacity. The results show that embedding the bottleneck as a portable prior in the decentralized strategy and guiding diversion with reward signals ahead of the safety upper limit can achieve joint improvement in efficiency and safety at a controllable computational cost and is more reliable under structural variation conditions.

Keywords

Multi-agent reinforcement learning; Emergency evacuation; Bottleneck perception; State encoding; Reward decomposition

Received: 27 Aug 2026; Revised: 19 Sep 2026; Accepted: 24 Oct 2026; Published: 30 Oct 2026
Copyright: ©2026 The Author(s). Published by Digital Intelligence Press Limited. This is an open access article under the CC BY 4.0 license.

1. INTRODUCTION

Emergency evacuation in university campuses faces a structural difficulty distinct from single-topology scenarios such as subway stations and sports venues: the corridors and stairways in the teaching areas form a sequential bottleneck, and the high-density living in the dormitory areas results in an extremely uneven initial distribution. Public spaces and evacuation routes overlap in time and

space, and together they constitute an evacuation environment of "multi-layer heterogeneous, sparse bottlenecks but concentrated consequences". In such environments, there is a significant spatial dislocation between individual local movement decisions and the overall congestion consequences [1],[2]. When a large number of individuals act based on the density gradient of their proximity, they tend to converge towards the same bottleneck simultaneously, forming local-optimal-driven global congestion. In recent years, multi-agent reinforcement learning has shown the potential to migrate from fixed rules to adaptive strategies in group evacuation simulations. The centralized training - decentralized execution paradigm makes it possible for individuals to acquire coordinated behavior under only local observations [3],[4]. However, when such methods are applied to university complex spaces, two technical obstacles have not been systematically addressed: first, pure local observations cannot provide the intelligent agents with predictive information about downstream bottlenecks, and the bottlenecks are precisely the nodes where the consequences of evacuation decisions are most concentrated; second, the sparse team-level rewards cannot distinguish "delays caused by reasonable detours" from "congestion caused by blindly rushing into the bottleneck", and the policy gradient is ambiguous, and the learned behaviors tend to maximize short-term progress rather than global diversion.

Regarding these obstacles, existing research has been conducted along three paths, but all have identifiable gaps. The static path planning method based on graph search has interpretability and determinism when the topology is known, but its path allocation is fixed before the simulation begins and cannot respond to the dynamic evolution of bottlenecks within the rounds [5],[6]. Optimization methods simulate congestion feedback using ant colony pheromones and follow the paradigm of "centralized optimization before rounds, static execution within rounds". The social force and cellular automata models based on microsimulation can depict the details of individual avoidance and congestion propagation, but they are difficult to support path optimization at the strategic level, and the computational cost increases sharply with the scale [7],[8],[9]. In learning methods, single-agent reinforcement learning treats evacuating individuals as independent decision-makers, ignoring the strategic coupling between individuals at the bottleneck; standard multi-agent reinforcement learning introduces coordination mechanisms, but its state encoding usually uses MLP or global attention, the former has a limited receptive field due to the local neighborhood, and the latter faces a quadratic computational burden in large-scale scenarios [10],[11]. More critically, the reward design of these learning methods almost all follow a single form of negative evacuation time or its increment, and the sparsity in the early training and the ambiguity in attribution in the later stage are not treated as independent issues.

The contribution of this paper is precisely based on this gap. Unlike previous works which view bottlenecks as passive physical phenomena in the simulation output, this paper embeds the bottleneck graph as a portable structural prior into the state representation of the agent, enabling individuals during decentralized execution to still determine the strength of the correlation between their behaviors and each bottleneck; unlike previous works which view reward design as training details, this paper decomposes rewards and elevates it to a methodological component of the same status as state encoding, enabling bottleneck pressure to receive gradient signals at the reward level earlier than the safety upper bound. The two designs are coupled in the value decomposition of centralized training, and their combined effect is not a simple superposition of perception and motivation, but a synergy resulting from the alignment of "seeing the correct object" and "motivating the correct behavior" on the same structure. The empirical part of this paper uses a three-level campus scenario as the carrier, with a structured perturbation protocol and effect size verification as the evaluation framework, to test the performance retention ability of the above designs under benchmark conditions and structural variation conditions

and provides mechanism evidence for the early triggering timing of diversion at the behavioral level. The subsequent sections successively elaborate on the algorithm architecture, experimental system, result analysis and discussion, and the conclusion section summarizes the contribution boundaries and the directions for generalization.

2. ALGORITHM ARCHITECTURE DESIGN AND IMPROVEMENT

2.1 Formalization of the problem and symbolic system

The evacuation of universities is modeled as a partially observable Markov game, which is characterized by the multi-dimensional group $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{O}_i\}, \{\mathcal{A}_i\}, \mathcal{P}, \{r_i\}, \gamma, \mathcal{G} \rangle$. Among them, $\mathcal{N} = \{1, \ldots, N\}$ represents the set of agents (individual evacuees), N varies in the range of hundreds to thousands depending on the scale of the scenario; $\mathcal{S}$ represents the global state space, including the positions, velocities of all individuals and the occupancy counts of each functional area; $\mathcal{O}_i$ is the local observation space of agent i, $\mathcal{A}_i$ is its action space; $\mathcal{P}: \mathcal{S} \times \prod_i \mathcal{A}_i \to \Delta(\mathcal{S})$ is the state transition kernel; r_i is the immediate reward for the individual, $\gamma \in (0, 1)$ is the discount factor; $\mathcal{G}$ is the bottleneck graph defined later, serving as a structural prior shared by all agents. The observation of agent i at time t consists of three parts:

$$o_i^t = (\, \phi_{\text{loc}}(i,t),\, \phi_{\text{den}}(i,t),\, \phi_{\text{nbr}}(i,t)\,) \tag{1}$$

Among them, the ϕ_{loc} code represents the function area number where it is located and the topological distance to the nearest exit. ϕ_{den} is the occupancy rate vector of the node and its one-hop neighborhood. ϕ_{nbr} is the relative orientation and distance of up to K neighboring individuals within the neighborhood truncation radius. The action space adopts a discrete form:

$$\mathcal{A}_i = \{\text{Up/Down/Left/Right/Wait}\} \cup \{\text{Migrate along the designated bottleneck edge}\} \tag{2}$$

Where the direction action corresponds to the movement within the gratified function area, the migration action corresponds to crossing the function area boundary and entering the adjacent node, and the waiting action allows the agent to actively yield at the congestion critical point. The strategy $\pi_i(a_i; \mid o_i^t \theta)$ is executed decentralizedly and only relies on o_i^t and the pre-stored bottleneck graph $\mathcal{G}$, without accessing the global state.

The bottleneck graph is defined as a weighted directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$. The node set $\mathcal{V} = \mathcal{V}_{\text{room}} \cup \mathcal{V}_{\text{corr}} \cup \mathcal{V}_{\text{stair}} \cup \mathcal{V}_{\text{exit}}$ corresponds to four functional areas: classrooms, corridor sections, stairwells, and exits. The edge set $\mathcal{E}$ represents the physical connectivity relationships between these functional areas. The edge weights are encoded by combining the traffic capacity and historical congestion probability:

$$w_{uv} = \alpha \cdot \frac{c_{uv}}{c_{\max}} + (1 - \alpha) \cdot \hat{p}_{uv}^{\text{cong}},\ (u, v) \in \mathcal{E} \tag{3}$$

In the formula, c_{uv} represents the unit-time traffic capacity of edge (u, v) (people/second), $c_{\max}$ is the upper bound of the total graph capacity used for normalization, $\hat{p}_{uv}^{\text{cong}}$ is the empirical frequency of this edge entering a high-density state in historical simulations, and $\alpha \in [0, 1]$ is the weight coefficient of the trade-off between capacity and congestion. Nodes with a node betweenness centrality higher than the quantile threshold are labeled as bottleneck nodes, forming the bottleneck node set $\mathcal{V}_b \subseteq \mathcal{V}$, whose size $M = \mid \mathcal{V}_b \mid$ is usually much smaller than $\mid \mathcal{V} \mid$. This sparsity constitutes

the structural basis for subsequent complexity optimization. Table 1 presents the key parameter configurations of the bottleneck graph in typical composite area scenarios to illustrate the controllability of the graph structure in terms of scale.

Table 1. Configuration of structural parameters of the bottleneck diagram (teaching-housing composite area scenario)

Parameters	Symbol	Value	Explanation
Total number of functional area nodes	$\lvert \mathcal{V} \rvert$	186	Including classrooms, corridors, stairs and exits
Number of bottleneck nodes	M	23	Top 12.4% by betweenness centrality
Number of connected edges	$\lvert \mathcal{E} \rvert$	412	Bidirectional edges are regarded as two directed edges.
Capacity trade-off coefficient	α	0.6	Capacity-oriented, congestion-anticipation-assisted
Neighborhood truncation scale	K	8	Number of the most recent individuals retained during the observation
Bottleneck embedding dimension	d	64	Bottleneck token vector dimension

As shown in Table 1, the bottleneck nodes only account for approximately one-eighth of the total number of nodes. The bottleneck graph compresses the scale of the nodes that need to be aggregated to a manageable range while maintaining the global structural information.

The objective of optimization of the strategy is to minimize the group evacuation time while maintaining the peak density within the bottleneck area. This can be expressed in a constrained expectation form:

$$\min_{\theta} \mathbb{E}_{\pi_\theta}\left[\max_{i\in\mathcal{N}} T_i^{\text{exit}}\right] \text{ s.t. } \max_{t}\max_{v\in\mathcal{V}_b}\rho_v^t \le \rho_{\text{safe}} \tag{4}$$

Among them, T_i^{exit} represents the completion time of the evacuation for individual i, ρ_v^t represents the occupancy rate of node v at time t, and ρ_{safe} represents the upper limit of safety density. This objective measures efficiency by the maximum value of the group evacuation time (i.e., the time taken by the last individual to evacuate), avoiding the concealment of tail delays by mean indicators, and simultaneously limits the density peak of bottleneck nodes in a hard constraint form, achieving an explicit trade-off between efficiency and safety at the optimization level.

2.2 Topology-aware hierarchical state encoder (TAHSE)

The inherent defect of purely local observation lies in the spatial dislocation between the information field and the decision outcome: Agents can perceive the congestion of neighboring individuals, but they cannot foresee the bottlenecks that are forming beyond several functional areas. The consequence is that a large number of individuals, driven by the local density gradient, simultaneously rush towards the same bottleneck, resulting in a typical failure mode of "local optimality - global congestion". The mechanism of this failure mode is that local observation only provides a first-order gradient of density, while the formation of a bottleneck is a second-order cumulative effect of density under topological constraints. Only first-order information cannot trigger diversion in advance in terms of time. The design starting point of TAHSE lies in: without expanding the observation space

or introducing communication, each agent carries a compressed priority about the global distribution of bottlenecks, thereby transforming "post-event passive response" into "pre-event active avoidance".

TAHSE employs a two-layer graph attention structure. The first layer performs aggregation on the local neighborhood subgraph, forming a subgraph $\mathcal{G}_i^{\text{loc}}$ consisting of the current node of the agent i and its one-hop neighborhood. The node features are composed of the occupancy rate, capacity, and distance to the exit of this functional area. The agent's own features and those of up to K neighboring individuals are used as queries, and are aggregated into a local context vector through the standard attention mechanism [12],[13]:

$$h_i^{(1)} = \sum_{j \in \mathcal{N}_K(i) \cup \{i\}} \text{softmax}\left(\frac{(W_Q x_i)^\top (W_K x_j)}{\sqrt{d_k}}\right) W_V x_j \tag{5}$$

In the formula, x_i represents the input features of the intelligent agent i, $\mathcal{N}_K(i)$ is the set of its K nearest neighbors, W_Q, W_K, W_V is the learnable projection matrix, and d_k is the attention key dimension. The output of the first layer, $h_i^{(1)}$, only contains local field information. Its receptive field is constrained by K and still cannot cover the distant bottleneck.

The second layer is the core of TAHSE: a set of learnable bottleneck tokens $\{e_m\}_{m=1}^{M}$ is introduced. Each token corresponds to a bottleneck node in the bottleneck graph. The initial embedding of the token is obtained by linearly mapping the graph structure features (betweenness centrality, capacity, mean of adjacent edge weights) of the node, and is jointly optimized with the policy parameters during training. The second layer uses $h_i^{(1)}$ as the query and all the bottleneck tokens as the key-value pairs to perform cross-domain attention, projecting local features to the global bottleneck representation space:

$$z_i = \sum_{m=1}^{M} \text{softmax}\left(\frac{\left(W_Q' h_i^{(1)}\right)^\top (W_K' e_m)}{\sqrt{d}}\right) W_V' e_m \tag{6}$$

Among them, $z_i \in \mathbb{R}^d$ is the bottleneck perception vector of the intelligent agent i, d is the bottleneck embedding dimension, and W_Q', W_K', W_V' is the second-layer projection matrix. The final state is represented as the concatenation of the local context and the bottleneck perception vector, $s_i = \left[h_i^{(1)}; z_i\right]$, which is then sent to the Actor network to output the action distribution. The essence of this design lies in: the bottleneck token acts as an "information bridge" between the global structure and the local observation. The intelligent agent does not need to observe the remote state; it only needs to judge the strength of the correlation between its local situation and each bottleneck through the attention weight and can obtain the implicit judgment of "my current behavior will exacerbate which bottleneck". The attention weight itself is also interpretable and can be used as a diagnostic basis for policy behavior.

In terms of complexity, the full graph attention needs to establish attention between all N intelligent agents and all $|\mathcal{V}|$ nodes. The complexity of a single layer is $O(N^2 + N|\mathcal{V}|)$. The first layer of TAHSE limits aggregation to the K neighbors, with a complexity of $O(N \cdot K \cdot d_k)$. The second layer only performs attention on M bottleneck tokens, with a complexity of $O(N \cdot M \cdot d)$. The total complexity is reduced from $O(N^2)$ to $O(N \cdot K + M \cdot d)$. Taking the parameters in Table 1 as an example, $K = 8$, $M = 23$, and $d = 64$. When $N = 1500$, the computational cost of the second layer attention is approximately 1.5% of the full graph scheme, and it grows linearly rather than

quadratically with N, which enables TAHSE to still meet the real-time reasoning requirements in large-scale global scenarios. Table 2 compares the computational costs and information coverage characteristics of different encoding schemes.

Table 2. Comparison of computational complexity and information coverage of state coding schemes

Coding scheme	Single-step complexity	Global bottleneck information	Interpretability	Applicable scale
MLP local encoding	$O(N \cdot d)$	No	Weak	Any scale
Global attention	$O(N^2 + N \mid \mathcal{V} \mid)$	Complete	Medium	Small and medium-sized scale
TAHSE (Proposed)	$O(N \cdot K + M \cdot d)$	Bottleneck-level compression	Strength (attention weight)	Large-scale

TAHSE replaces the complete global coverage with bottleneck-level compression. What it sacrifices is the detailed information in non-bottleneck areas, while what it gains is a leap in complexity from quadratic level to linear level. Considering that the decision consequences during the evacuation process mainly concentrate on bottleneck nodes, this information trade-off achieves a balance that is beneficial for large-scale deployment between efficiency and security.

2.3 Bottleneck sensitive reward decomposition mechanism (BSRD)

The standard team rewards typically take the negative group evacuation time or its increment, and they have a dual nature. Firstly, sparsity: Before the evacuation is completed, the team rewards hardly produce an effective gradient, and the early training strategies are almost randomly wandering in a huge action space. Secondly, attribution ambiguity: Negative time rewards cannot distinguish "delays caused by reasonable detours" from "traffic congestion delays caused by blindly rushing into bottlenecks", and these two types of behaviors cannot be differentiated at the reward level, resulting in ambiguous gradient directions and the learned strategies tending to maximize short-term progress rather than achieve global diversion. The design goal of BSRD is to decompose a single team signal into three clearly directional components, so that "moving towards the exit", "avoiding bottleneck pressure", and "completing area clearance" three types of behaviors each receive identifiable gradient sources.

The immediate team rewards after decomposition are written as a weighted sum of three items:

$$r^t = \lambda_1 r^t_{\text{prog}} + \lambda_2 r^t_{\text{press}} + \lambda_3 r^t_{\text{clr}} \tag{7}$$

Progress Reward r^t_{prog} measures the normalized reduction amount of the distance from all intelligent agents to the nearest exit topology, and its form is [14]:

$$r^t_{\text{prog}} = \frac{1}{N} \sum_{i \in \mathcal{N}} \frac{d_{\mathcal{G}}\left(v_i^{t-1}, v^*_{\text{exit}}\right) - d_{\mathcal{G}}\left(v_i^t, v^*_{\text{exit}}\right)}{d_{\mathcal{G}}\left(v_i^0, v^*_{\text{exit}}\right)} \tag{8}$$

In the formula, $d_{\mathcal{G}}(\cdot,\cdot)$ represents the shortest topological distance on the bottleneck graph, v_i^t is the node, where the agent i is located at time t, v^*_{exit} is its nearest exit node, and the denominator is the initial distance used for normalization to make the contributions of individuals from different initial

positions comparable. The bottleneck pressure penalty r_{press}^{t} calculates the cost for an individual to enter the high-density area based on the marginal increment of the real-time occupancy rate of the bottleneck node:

$$r_{\text{press}}^{t} = -\frac{1}{|\mathcal{V}_b|}\sum_{v\in\mathcal{V}_b} \max\ (0, \rho_v^t - \rho_v^{t-1}) \cdot \mathbb{1}[\rho_v^t > \rho_{\text{warn}}] \tag{9}$$

Among them, ρ_{warn} is the bottleneck warning threshold (lower than the safety upper limit ρ_{safe}), and the indicator function ensures that the penalty only takes effect after the occupancy rate exceeds the warning line, avoiding excessive suppression of normal traffic in the early stage of training; the existence of this threshold makes the penalty term have a "lead time" in terms of time. The reward r_{clr}^{t} is coordinated to provide a phased positive signal when any functional area is emptied, with the amplitude proportional to the initial number of people in that area, to alleviate sparsity and provide a credit anchor for regional-level coordinated behavior.

The three coefficients, $\lambda_1, \lambda_2, \lambda_3$, are not fixed parameters but are determined through adaptive adjustment during the training process. The adjustment is based on the relative imbalance degree of the two diagnostic quantities within the rolling window: if the density of bottleneck peaks in the window continuously approaches ρ_{safe} while the improvement in evacuation time stagnates, then increase λ_2; if the variance of the progress reward is too low and the regional clearance events are rare, then increase λ_3. The adaptive law adopts a damp multiplicative update, and the update step size decreases with the number of training rounds to ensure stability in the later stages. Table 3 presents the typical coefficient configurations for each stage of the course, showing that the coefficients systematically tilt towards safety items as the difficulty of the scenario increases.

Table 3. Configuration of BSRD coefficients at each stage of the course

Training phase	λ_1 (Progress)	λ_2 (Pressure)	λ_3 (Clear)	Regulation basis
Phase one (low-density fixed output)	1.0	0.3	0.2	Progress-oriented
Phase two (random exit closure)	0.8	0.9	0.4	Increase in pressure
Phase three (distribution shift and dynamic obstacles)	0.6	1.4	0.6	Safety first

The coefficient configuration is not a pre-set value before training, but the steady-state value obtained after the adaptive law converges at each stage. The changing trend reflects the transfer of the weight of safety constraints relative to the efficiency target as the difficulty of the scene increases.

The integration with the CTDE paradigm follows a standard centralized training-decentralized execution structure, but the decomposition of the value function has characteristics unique to this paper. The centralized Critic receives the global bottleneck graph state G^t (the vector of the occupancy rates of all bottleneck nodes) and the concatenation of all agent observations and outputs a joint value estimate. Its decomposition form is:

$$Q^{\text{tot}}(G^t, o^t, a^t) = \underbrace{V_{\text{team}}(G^t)}_{\text{Overall bottleneck situation}} + \sum_{i\in\mathcal{N}} \underbrace{A_i(s_i^t, a_i^t, G^t)}_{\text{Individual advantage}} \tag{10}$$

Among them, V_{team} only relies on the global bottleneck situation to depict the team's baseline value that is independent of individual actions; A_i is the individual advantage function, with the mixed representation s_i^t, action a_i^t, and global bottleneck situation as the input. The essence of this decomposition is to mainly attribute the variance of team value to the global term, and the individual advantage only explains the differences in actions, so that in the execution stage, the Actor only needs s_i^t to approximate the ranking of individual advantages, without accessing G^t. The three rewards of BSRD enter the temporal difference objective of the Critic according to their weights, so that the credit allocation of each component remains separable at the value level, avoiding the mutual cancellation of the three in a single temporal difference error.

3. SIMULATION EXPERIMENT SYSTEM AND VERIFICATION METHODS

3.1 Construction of simulation environment and abstraction of campus scenarios

The simulation environment is customized based on the Petting Zoo multi-agent interface [15],[16]. The underlying representation is a hybrid of discrete-time step grid and graph: the internal area of the functional zone is depicted with a fine grid to show the position of individual agents and local avoidance, while the areas between functional zones are connected by edge of the bottleneck graph. When an agent crosses the boundary of the functional zone, the movement action on the graph is used to complete the crossing. The time step is set to 0.25 seconds, which matches the typical time scale of individual reactions and movements in crowd evacuation. The termination condition for each round is that all agents reach the exit, or the simulation duration reaches the upper limit. The upper limit is set to 2.5 times the rated evacuation time for the scenario size, to prevent individual agents who are stuck from prolonging the round indefinitely. The environment supports parallel multi-round sampling. The simulation cost for a single scenario and a single round is approximately linearly related to the number of agents. This feature makes the multi-scale scenarios and the repeated use of multiple sub-scenarios described later computationally feasible.

The abstraction of the three-level scenarios follows the principle of "increasing structural complexity rather than merely increasing scale". The single teaching building scenario has approximately 500 agents, with a topology of a three-story structure of a single teaching building, where the uniqueness of the exits is high and the bottlenecks are concentrated in the stairwells, which is used to test the basic performance of the method in scenarios with sparse bottlenecks; the teaching-housing combined area scenario has approximately 1500 agents, with a topology consisting of the teaching area, the housing area, and the common passage connecting the two, with multiple exits and multiple path options, and the bottlenecks present cross-regional coupling, which is the most valuable scenario for testing the predictive ability of the method; the university-wide scenario has approximately 3000 agents, with additional public spaces such as the cafeteria and the gymnasium added on top of the combined area, and more redundant paths, which is used to test the scalability and computational feasibility of the method under scale growth. Table 4 presents the structural parameter configurations of the three scenarios, and each parameter is a component of the scenario definition rather than an experimental observation result.

Table 4. Configuration parameters of three-level campus scenario structure

Parameters	Single teaching building	Teaching and dormitory combined area	Entire school area
Rated number of intelligent agents	500	1500	3000
Number of nodes in the functional area	64	186	402
Bottleneck node count: M	9	23	51
Number of exits	2	6	11
Average path length (number of edges)	7.2	11.8	15.4
Narrowest bottleneck edge capacity (persons/second)	1.6	1.2	1.2

The number of bottleneck nodes increases with the total number of nodes but their proportion decreases (14.1% for a single teaching building, 12.4% for a compound area, and 12.7% for the entire campus area). This indicates that the bottleneck graph remains sparse across different scales. The narrowest bottleneck edge capacity of the compound area and the entire campus area is the same, while the average path length increases from 11.8 to 15.4, suggesting that the bottleneck pressure in the entire campus scenario is more due to the diversion decisions resulting from path redundancy rather than the capacity of a single channel.

Scene parameterization is achieved through a unified scene description file, which contains five fields: the node-edge structure of the topology graph, the capacity of each node, the exit position and width, and the initial distribution of intelligent agents. Any modification of any field will generate a scene variant, and the perturbation protocol is applied to this parameterization interface with structured modifications. Bottleneck annotation adopts dual-source fusion: one source is the empirical frequency of nodes entering a high-density state in historical congestion simulations, and the other source is the topological betweenness centrality. Nodes that exceed their respective quantile thresholds in both sources are included in the bottleneck node set. The fusion rule can be written as the bottleneck criterion for node v:

$$\mathbb{1}[v \in \mathcal{V}_b] = \mathbb{1}\left[\hat{p}_v^{\text{cong}} \geq q_{\text{cong}}\right] \cdot \mathbb{1}[B_v \geq q_{\text{betw}}] \tag{11}$$

In the formula, $\hat{p}_v^{\text{cong}}$ represents the empirical frequency at which node v enters the high-density state, B_v is its betweenness centrality, and q_{cong} and q_{betw} are the quantile thresholds for the two sources.

3.2 Comparison of baseline settings

The baseline system is configured in four layers: "rules - optimization - learning - upper bound". Each layer performs different comparison functions. The rule-based baseline involves Dijkstra's shortest path routing combined with local density avoidance rules [17],[18]. Individuals travel along the topological shortest path to the nearest exit. When the adjacent density exceeds the threshold, lateral avoidance or short-term waiting is executed. This baseline represents the mainstream technical form of the current campus evacuation plan, and its value lies in providing a performance lower bound reference under the condition of "no learning and no global coordination". The optimization baseline is the

adapted version of behavior-aware adaptive ant colony optimization in the campus scenario. It uses ant colony pheromones to simulate path congestion feedback and uses behavior-aware factors to regulate the randomness of individual path selection. The path allocation plan is calculated once at the beginning of each round and is not updated within the round. This baseline represents the current level of group intelligence methods in the path planning aspect. Its paradigm of "centralized optimization before rounds, static execution within rounds" forms a methodological contrast with the "dynamic response during execution" paradigm in this paper.

The optimization of the path allocation of the baseline can be formalized as a two-stage process consisting of pheromone update and path selection. The pheromone update is:

$$\tau_{uv} \leftarrow (1-\rho)\tau_{uv} + \rho \sum_{k=1}^{K_{\text{ant}}} \Delta\tau_{uv}^{(k)},\ \Delta\tau_{uv}^{(k)} = \frac{Q}{L_k} \tag{12}$$

In the formula, τ_{uv} represents the pheromone concentration on edge (u, v), ρ represents the evaporation rate, K_{ant} represents the ant colony size, Q represents the pheromone intensity constant, and L_k represents the congestion-weighted length of the path constructed by the kth ant. The path selection probability is obtained by normalizing the product of pheromone and behavioral knowledge factors, and the behavioral knowledge factor is adjusted based on the real-time occupancy rate of nodes to regulate the randomness of selection. The key point of this formula is that the update is only executed once at the beginning of each round, and τ_{uv} is frozen within the round. This is precisely the mathematical root reason why it cannot respond to the dynamic evolution of bottlenecks within the round.

The learning baseline is the standard MAPPO, where both the Actor and Critic adopt MLP local encoding, and the reward adopts a single negative evacuation time team reward [19],[20]. Apart from the two architectural contributions of this paper, all the training configurations (course settings, hyperparameters, interaction budget) are exactly the same as those of the method in this paper. The function of this baseline is to isolate the respective gains of TAHSE and BSRD: The differences between this method and MAPPO can be fully attributed to the two architectural contributions, while the separate ablation of TAHSE and BSRD and the comparison chain of MAPPO together constitute, further, decomposes the total gain into the independent effects and interaction effects of the two contributions [21]. The upper bound reference is the idealized strategy that provides global bottleneck information and centralized scheduling. It accesses the global state at each decision step and performs centralized diversion, only used to indicate the performance upper bound and not participating in the fairness ranking. It is separately listed in the report and not included in the significance test. The four-layer baseline progresses from local to global in information conditions and from static to dynamic in decision paradigms. The method in this paper is in the "local observation + decentralized execution" grid, and the upper bound reference is in the "global state + centralized scheduling" grid. The performance gap between them measures "the degree of approaching the centralized upper bound under decentralized constraints".

3.3 Evaluation index system

The indicator system is organized into four categories: efficiency, safety, robustness, and training efficiency. Each category of indicators clearly defines its statistical object and reporting format. The efficiency indicators mainly rely on the 95th percentile of the total evacuation time, with the average individual evacuation time as a secondary measure. The reason for using the 95th percentile instead of

the mean is that the tail delays in the evacuation process are dominated by a few individuals who are stuck, and the mean would be diluted by a large number of early evacuees, failing to reflect the critical safety metric of "when the last group of people is out of danger"; the 95th percentile is statistically stable and semantically corresponds to "almost all evacuations" at a certain moment. Let the evacuation time set of all individuals in the k repetition be $\{T_i^{\text{exit}}\}_{i\in\mathcal{N}}$. Then the efficiency indicators are written as:

$$T^{95\%} = \text{Quantile}_{0.95}\left(\{T_i^{\text{exit}}\}_{i\in\mathcal{N}}\right),\ \bar{T} = \frac{1}{N}\sum_{i\in\mathcal{N}} T_i^{\text{exit}} \tag{13}$$

$\text{Quantile}_{0.95}(\cdot)$ represents the 95% quantile function of the sample, and $\bar{T}$ represents the average individual evacuation time. The reporting method that uses the quantile as the main indicator and the mean as the secondary indicator enables both tail delays and overall efficiency to be observed within the same framework, avoiding the concealment of individual differences by a single mean value.

The safety indicators include three items: peak density of bottleneck nodes, duration of high-density state, and spread range of congestion. These three indicators respectively depict the instantaneous extreme value of density, the cumulative duration of the over-threshold state, and the extent of congestion spreading in the topology. The three indicators can be collectively expressed as:

$$\rho^{\text{peak}} = \max_t \max_{v\in\mathcal{V}_b} \rho_v^t,\ D^{\text{high}} = \sum_t \mathbb{1}\left[\max_{v\in\mathcal{V}_b} \rho_v^t > \rho_{\text{safe}}\right]\Delta t,\ S^{\text{cong}} = \frac{|\{v, \in \mathcal{V}: \exists t \rho_v^t > \rho_{\text{safe}}\}|}{|\mathcal{V}|} \tag{14}$$

In the formula, ρ_v^t represents the occupancy rate of node v at time t. Δt represents the time step, and S^{cong} represents the proportion of nodes that have entered the super-threshold state out of all nodes in the graph. These three indicators correspond to the three safety dimensions of "worst instant", "cumulative exposure", and "spatial diffusion". The degradation in any one dimension cannot be compensated for by the other two. This multi-dimensional structure ensures that the safety assessment is not dominated by a single peak indicator.

The robustness index is defined as the rate of performance retention on the perturbation scenario set. Its general form is the normalized degradation amplitude of the index value under perturbation relative to the index value of the benchmark scenario [22]:

$$R_k = 1 - \frac{|\mu_k^{\text{pert}} - \mu_k^{\text{base}}|}{\mu_k^{\text{base}}} \tag{15}$$

In the formula, μ_k^{pert} represents the mean value of the indicator in the perturbation scenario set for the item k, while μ_k^{base} represents the mean value of the same indicator in the unperturbed reference scenario. The value of R_k is closer to 1, indicating better performance retention. For indicators that are "the smaller, the better" (such as evacuation time and peak density), this formula is directly applicable. For indicators that are "the larger, the better", the opposite sign is taken in the numerator. The retention rate is calculated separately for each perturbation dimension, forming a retention rate matrix of "method × perturbation dimension", enabling the robustness conclusion to be attributed to specific perturbation types rather than a general "robust" assertion. The training efficiency indicators include the number of interaction steps required to achieve the target performance and the number of convergences rounds in each stage of the course. The target performance threshold is set at 90% of the final performance of the learning baseline to measure the relative cost of the method in terms of training cost. The number of

interaction steps is reported as the median to suppress the interference of long-tail differences between seeds in the cost comparison.

3.4 Structured perturbation protocol and statistical test scheme

The disturbance protocol is systematically applied along three dimensions, each dimension corresponding to a type of structural variation with a clear physical origin in the actual campus evacuation [23]. Topological disturbance simulates channel failure and capacity degradation. The specific operation is to randomly close a non-unique exit or reduce the capacity of a certain bottleneck edge by 30%. The former changes the set of feasible paths, while the latter changes the bottleneck strength without changing the connectivity. These two types of operations are used to test the adaptability of the strategy to "path re-planning" and "bottleneck redistribution". The capacity reduction operation can be expressed as edge weight disturbance:

$$c_{uv} \leftarrow c_{uv} \cdot (1-\delta),\ \delta = 0.30 \tag{16}$$

In the formula, c_{uv} represents the traffic capacity of the edge (u, v), and δ represents the capacity degradation ratio; the distribution perturbation simulates the unexpected deviation of the spatial distribution of the crowd at the start of the evacuation, shifting the initial agent distribution from the training distribution area to the non-training distribution area as a whole, with the shift amplitude set according to a fixed proportion of the scale of the functional area of the scenario, and the test strategy examines the stability under the observed distribution deviation. The scale perturbation simulates the fluctuation of the actual evacuation population relative to the rated capacity, changing the number of agents within ±40% of the rated value at fixed intervals, and the test strategy examines the performance retention under the change of density conditions.

The statistical test plan selects the test method based on the comparison type: for the comparison between the method in this paper and the baseline under the same scenario and the same perturbation variant, it is a paired comparison (sharing the scenario instance and perturbation instance), and the Wilcoxon signed-rank test is used; for the comparison across scenarios or involving different scenario instances, it is an unpaired comparison, and the Mann-Whitney U test is used. For each comparison group, at least 20 independent training seed repetitions are performed. The seed simultaneously controls the initialization strategy, the order of course sampling, and the sampling of perturbation instances, so that the sources of variation between repetitions can be traced. The test results also report the p-value and the effect size Cohen's d, and the effect size is calculated according to the following formula:

$$d = \frac{\bar{X}_{\text{ours}} - \bar{X}_{\text{base}}}{s_{\text{pooled}}},\ s_{\text{pooled}} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \tag{17}$$

$\bar{X}_{\text{ours}}$ and $\bar{X}_{\text{base}}$ represent the mean values of the indicators of the method in this paper and the baseline in the repeated experiments, s_{pooled} is the combined standard deviation, and n_1, n_2 is the number of repeated trials for the two groups. The conclusions are classified into two categories: "statistically significant but with limited practical significance" and "statistically and practically significant" based on the effect size threshold. The reason is as follows: Even small differences can reach statistical significance under large sample repetitions. If only judged by the p-value, the practical value of the method advantage is systematically overestimated; the size effect places the differences on the standard deviation scale, making "whether the advantage is worth deployment" a determinable issue. The significance level is set at 0.05, and multiple comparison corrections are made. The correction

method is explicitly stated in the report.

The ablation design has four variants: TAHSE removal (replacing with MLP local encoding), BSRD removal (replacing with a single negative evacuation time team reward), simultaneous removal of both (equivalent to the standard MAPPO), and the complete method. The four variants perform the same repetition and verification process as the main experiment on all perturbation scenarios, thereby evaluating the independent and interactive effects of the two architectures within the same statistical framework. The determination of the interactive effect does not rely on subjective interpretation but is characterized by the additive deviation quantity:

$$\Delta_{\text{inter}} = (\bar{X}_{\text{full}} - \bar{X}_{\text{both}}) - [(\bar{X}_{\text{noTAHSE}} - \bar{X}_{\text{both}}) + (\bar{X}_{\text{noBSRD}} - \bar{X}_{\text{both}})] \tag{18}$$

$\bar{X}_{\text{full}}$ represents the mean value of the complete method indicators, $\bar{X}_{\text{both}}$ represents the mean value when both variants are removed simultaneously, $\bar{X}_{\text{noTAHSE}}$ and $\bar{X}_{\text{noBSRD}}$ are the mean values of the individual ablated variants. Δ_{inter} significantly deviating from zero indicates the existence of non-additive interaction between the two contributions, and its sign indicates whether the interaction is synergistic or antagonistic.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Comparison of baseline scenario performance

The comparison of efficiency and safety indicators for three scale scenarios under undisturbed conditions is shown in Figure 1. Figure 1 adopts a double-panel structure: the left panel represents the 95th percentile of the total evacuation time, and the right panel represents the peak density of bottleneck nodes. Both panels are grouped by scenarios and have bars within each group, and the error bars represent the standard deviation of 20 seed repetitions. Overall, in the composite area and the entire domain scenarios, the method proposed in this paper simultaneously achieves dual advantages in efficiency and safety. However, in the single teaching building scenario, the efficiency advantage has significantly narrowed, while the safety advantage remains. This uneven distribution of gains itself constitutes a direct test of the design motivation of TAHSE.

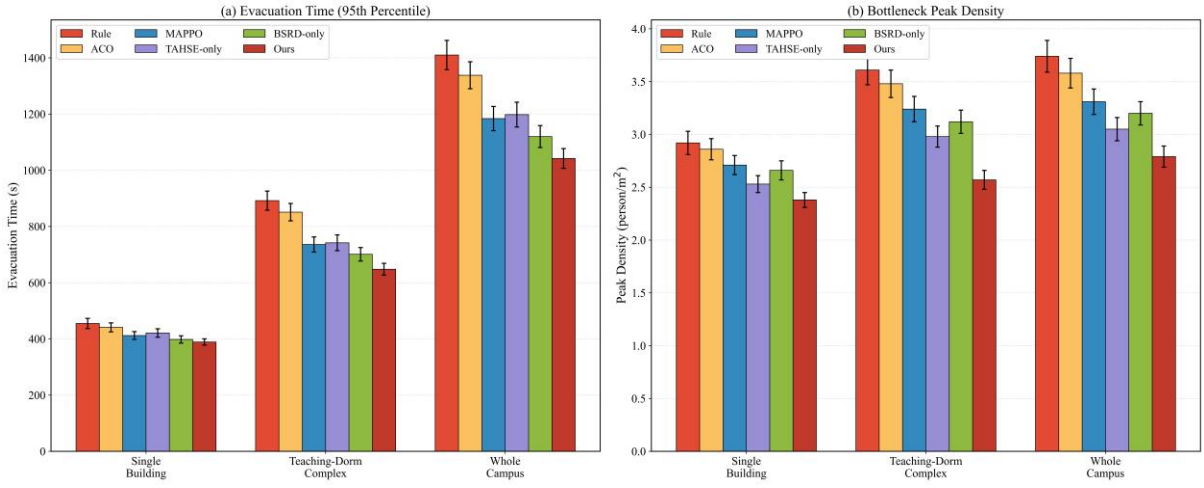


Figure 1. Comparison of efficiency and safety indicators under three scenarios without disturbance

The gain amplitude of the composite area scenario is the largest. The peak density reduction of the proposed method relative to the MAPPO baseline is significantly greater than that of the single teaching building scenario. The bottleneck node number of the composite area is 23, the exit number is 6, and the average path length is 11.8. Both the bottleneck quantity and the path redundancy are at a medium-high level. The intelligent agent has a real choice space among multiple feasible paths, and local observation cannot distinguish the bottleneck states downstream of these paths. It is precisely the combination of "having a choice space but no foresight ability" that enables the bottleneck token of TAHSE to achieve the maximum value. The gain narrowing of the single teaching building scenario is due to the simplification of the bottleneck: there are only 9 bottleneck nodes, 2 exits, and an average path length of 7.2. The feasible path set itself is narrow, even if the intelligent agent accurately predicts the bottleneck in the stairwell, the available detour schemes are extremely limited. The predictive value is compressed by topological constraints. The absolute gain of the all-school area is higher than that of the single teaching building but lower than that of the composite area. The reason is that the bottleneck pressure in the all-school area mainly comes from the diversion decision under path redundancy rather than the capacity of a single channel. TAHSE mainly plays the role of "balancing the load among multiple redundant paths" in this area, and the form of gain is different from that of the composite area's "pre-emptively avoiding key bottlenecks".

The comparison with the two non-learning baselines reveals the source structure of the gain. The rule baseline and the optimization baseline have acceptable safety indicators in the single teaching building scenario, but in the composite area and the entire school area scenarios, the peak density has significantly increased. Among them, the "pre-round centralized optimization and static execution within the round" paradigm of the optimization baseline fails particularly obviously during the distribution dynamic evolution: its path allocation has been fixed at the beginning of the round, and when the leading individual triggers an early saturation of a certain bottleneck, it is unable to re-allocate the subsequent individuals. This failure mode contrasts sharply with the "dynamic response during execution" paradigm of the method in this paper. The remaining gap between the method in this paper and the upper bound reference exists in all three scenarios but has a limited amplitude, indicating that the performance loss under decentralized constraints has been controlled within a small range; this gap is slightly larger in the entire school area scenario, suggesting that the marginal difficulty of scale growth on decentralized coordination is still increasing. **Table 5** summarizes the mean values of the indicators corresponding to each method in the three scenarios and the ratio of the gap to the upper bound reference as calculated relative to the method in this paper. The values in the table are the averages of 20 seeds, and the gap ratio is calculated based on the method in this paper.

Table 5. Differences between the mean values of each method and the upper bounds under the three scenarios without disturbance

Scene	Method	Evacuation time 95th percentile (s)	Peak density (people/square meter)	Relative upper bound gap (%)
Single teaching building	MAPPO baseline	412	2.71	18.4
Single teaching building	Method presented in this article	389	2.38	11.2
Composite zone	MAPPO baseline	736	3.24	26.7
Composite zone	Method presented in this article	648	2.57	9.8
Entire school area	MAPPO baseline	1184	3.31	31.5
Entire school area	Method presented in this article	1042	2.79	14.6

The gap ratio between the method described in this paper and the upper bound reference was within 15% in all three scenarios. However, the gap ratio of the MAPPO baseline increased from 18.4% to 31.5% as the scale of the scenarios expanded. The gap between the two methods widened as the scale increased, indicating that the performance loss under decentralized constraints was effectively suppressed in the architecture of this paper. Meanwhile, the gap ratio of the composite zone was the lowest (9.8%), which was consistent with the aforementioned judgment of "maximum gain in the composite zone" in terms of numerical values. The two are mutually corroborative rather than repetitive statements.

4.2 Ablation experiment: contribution attribution analysis

Figure 2 presents the normalized performance of the four ablation variants in four categories of indicators using a grouped bar chart. The x-axis represents the four categories of indicators (95th percentile of evacuation time, bottleneck peak density, high-density duration, congestion propagation range), and the y-axis represents the normalized values relative to the complete method in this study (with 1.0 as the reference). Within each group, there are four ablation variants. The differences in the height of the bars visually illustrate the division of responsibilities: The ablation variant of TAHSE alone shows significantly lower performance in two safety indicators and relatively higher efficiency indicators. In contrast, the ablation variant of BSRD is the opposite, with a more significant decrease in efficiency indicators and better performance in safety indicators. The complete method is close to or reaches 1.0 in all four categories of indicators.

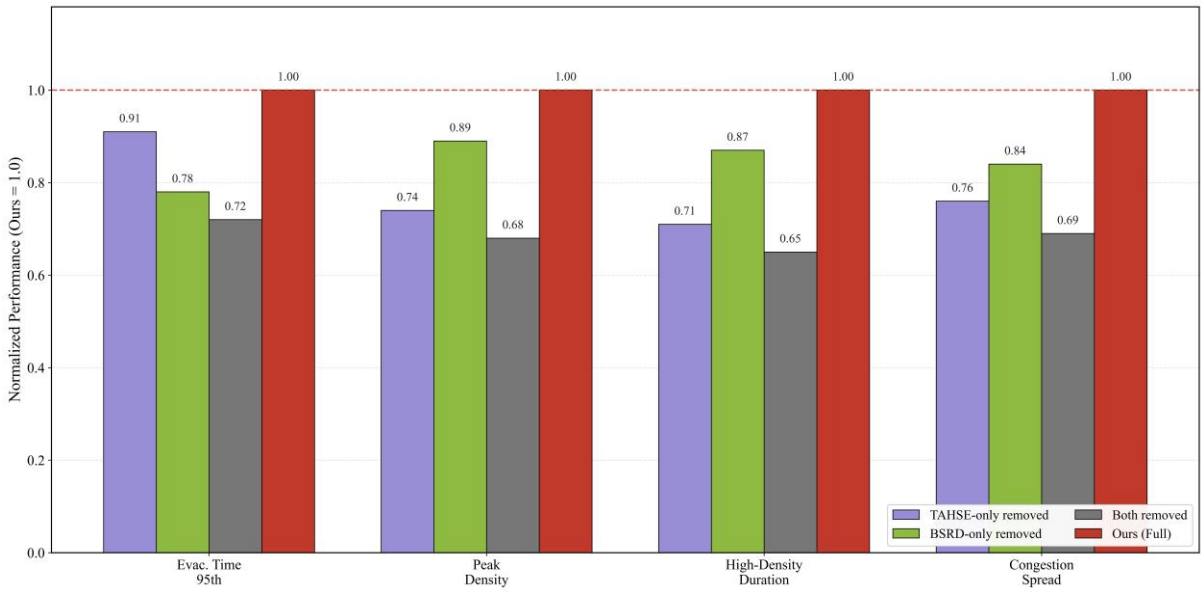


Figure 2. Comparison of normalized performance of the four types of ablation variants

TAHSE addresses the perception problem of "invisible bottlenecks", and its absence directly leads to the inability of the intelligent agent to predict downstream congestion, thus causing the degradation of safety indicators to the main level; BSRD addresses the signal problem of "seeing but without the incentive to respond", and its absence causes the strategy to degenerate into maximizing short-term progress, thus causing the tail delay of efficiency indicators to degenerate to the main level. The independence of the two indicates that perception and incentive are two orthogonal failure dimensions of the evacuation strategy, and the absence of either dimension cannot be compensated by the other.

When BSRD is used alone, the stability issue during the training phase is a phenomenon that needs to be diagnosed separately in the ablation analysis. Figure 3 presents the bottleneck peak density curves of the BSRD single-ablation variant and the complete method during the training process. The x-axis represents the interaction steps, and the y-axis represents the peak density within the rolling window. Both curves are presented as the median and quartile bands of 20 seeds. The BSRD single-ablation variant shows a sign of peak density recovery and an increase in variance between seeds in the later training stage. The diagnostic interpretation is as follows: when the bottleneck representation without TAHSE is lacking, the pressure penalty is triggered based on local density. The strategy gradually learns to "lower the local density reading" rather than "actually divert", that is, it responds speculatively to the penalty signal, resulting in the penalty term decoupling from the true safety target. In the complete method, the bottleneck representation of TAHSE anchors the penalty trigger condition to the global bottleneck occupancy rate, reducing the space for speculative responses, and the curve remains stable in the later stages.

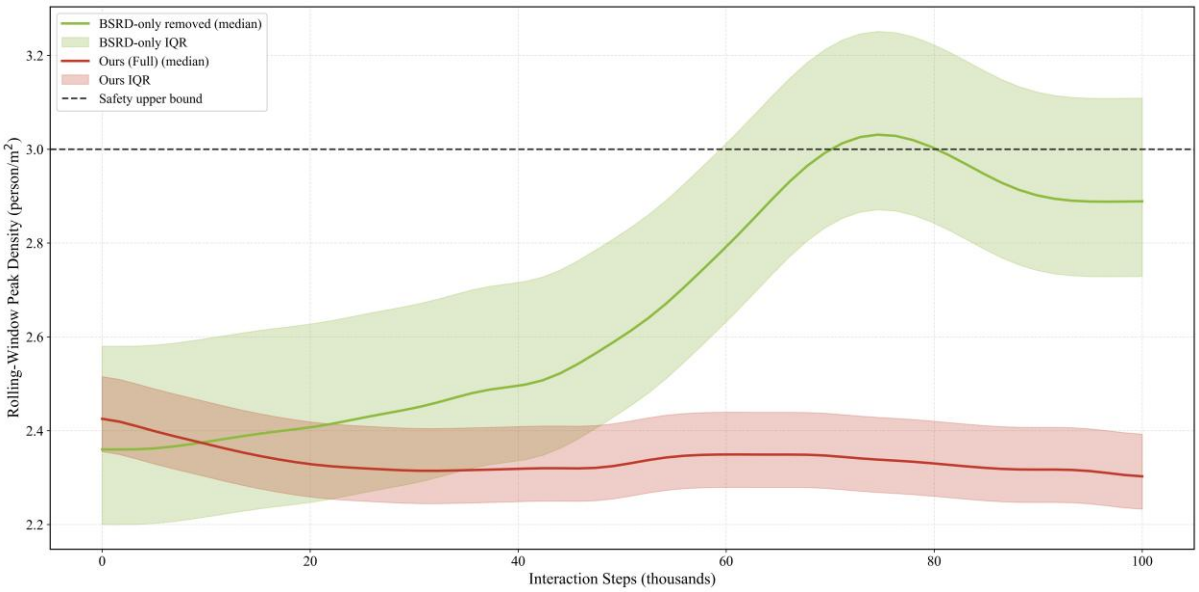


Figure 3. Peak density curve of training process bottleneck

When the diagnostic quantity detects that the peak density is approaching the safety upper limit and the efficiency improvement stalls, increasing the pressure coefficient enables the strategy to regain the shunt incentive. The interquartile range of the complete method curve narrows in the later stage, which is an indication of the effectiveness of this mechanism. To quantify this diagnostic, Table 6 lists the mean peak density, inter-variant standard deviation, and steady-state pressure coefficient values of each variant in the ablation experiment at the later stage (the last 20% interaction steps) of training. The inter-variant standard deviation of the BSRD single-ablation variant is significantly greater than that of the complete method, indicating that its instability in the later stage is not only a recovery at the mean level, but also an intensification of the differentiation of behaviors between seeds. The steady-state pressure coefficient value of the complete method under the adaptive law is higher than the initial configuration, confirming the corresponding relationship between the coefficient increase and the stability recovery.

Table 6. Diagnostic criteria for stability in the later stage of ablation variant training

Ablation variant	Peak density average (people/square meter)	Standard deviation among seeds	Steady-state value of pressure coefficient	Interquartile range width
TAHSE alone is ablated	2.94	0.21	0.31	0.38
BSRD alone ablation	3.12	0.34	0.28	0.57
Remove both at the same time	3.27	0.29	—	0.49
Complete method of this article	2.61	0.12	1.37	0.19

The standard deviation between seeds of the complete method and the inter-quartile range width are approximately one-third of those of the BSRD single ablation variant, while the steady-state value of the pressure coefficient is nearly five times that of it. These three values jointly point to the causal chain of "coefficient adaptive upregulation-penalty signal anchoring-stable training later"; although the steady-state value of the pressure coefficient of the TAHSE single ablation variant is slightly higher,

due to the lack of bottleneck representation, it cannot convert the penalty into effective diversion, and its peak density is still higher than that of the complete method. This comparison further demonstrates that perception and motivation are both indispensable.

4.3 Robustness test of disturbance

The robustness of perturbation is presented as a performance retention rate matrix. Figure 4 is a heatmap. The rows correspond to four methods (rule baseline, optimization baseline, MAPPO baseline, and the method in this paper), and the columns correspond to three perturbation dimensions and cross-dimension summaries. The color depth of the cells indicates the retention rate, and the numerical values are marked within the cells. Overall, the method in this paper has the highest retention rate in all four columns, and its advantage is most prominent in the topological perturbation column. Distribution perturbation and scale perturbation change the spatial distribution and quantity conditions of the crowd. The learning baseline has already acquired certain adaptability through course training, and the gap between the methods converges relatively; topological perturbation changes the feasible path set itself, requiring the strategy to redistribute the load after the bottleneck graph structure changes, which precisely is the design direction of the TAHSE bottleneck token mechanism.

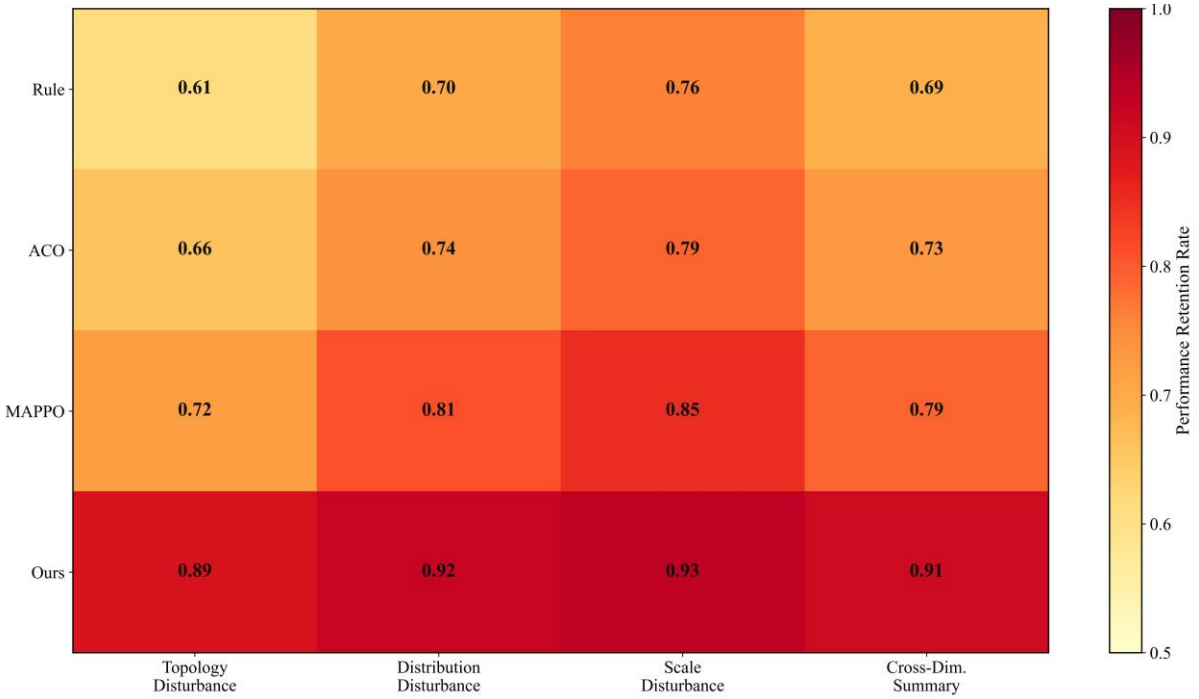


Figure 4. Heatmap of the performance retention rate of the disturbance scenario set

The performance under topological perturbations needs to be elaborated separately, as it is a key condition for testing the adaptability of TAHSE bottleneck tokens. Figure 5 focuses on the scenario of an exit being closed, presenting a grouped bar chart to compare the degradation amplitudes of the method in this paper and the MAPPO baseline in various security and efficiency indicators. The degradation amplitude is expressed as a percentage relative to the baseline scenario without perturbation. The shorter the bar, the smaller the degradation. The degradation amplitude of the method in all indicators is less than that of the learning baseline. Among them, the gap in the degradation of peak density and high-density duration is the most significant. The mechanism lies in: the closed exit changes the structure of the bottleneck node set. The MLP local encoding of the MAPPO baseline cannot

perceive this structural change. The policy still acts according to the local density-action mapping formed during training, causing individuals to continuously flow towards the upstream channel of the failed exit; after the bottleneck token of TAHSE is closed, its attention weight can be redistributed based on the pre-stored bottleneck graph's structural changes, enabling the agent to adjust its judgment of "which bottleneck is currently more critical" according to the topological changes.

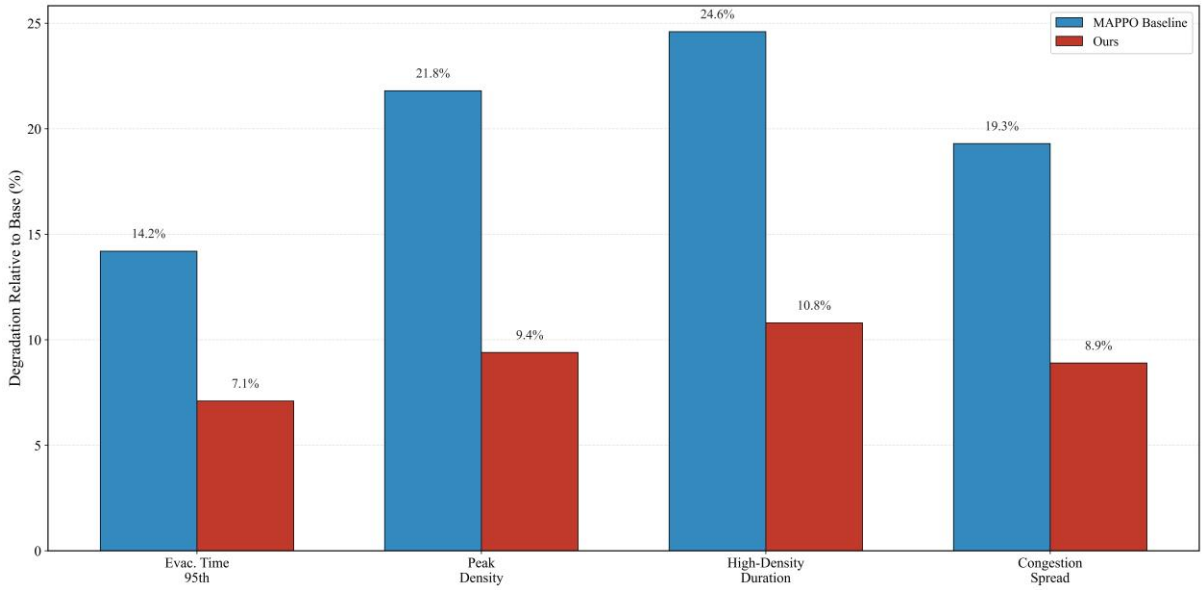


Figure 5. Comparison of degradation amplitudes of each indicator in the exit-closed scenario

The statistical test results are presented in Figure 6. This figure is a forest plot. The horizontal axis represents the effect size (Cohen's d) of the method in this paper relative to the MAPPO baseline on each indicator. The vertical axis represents the indicator's name. Each indicator corresponds to a point estimate and a confidence interval and is assigned to different marker zones to represent the results of the Wilcoxon test and the unpaired Mann-Whitney U test. In the figure, vertical reference lines indicate the practical significance threshold of the effect size, and indicators located to the right of the reference line indicate both statistical and practical significance. Indicators located within the statistical significance line but not reaching the practical threshold are presented with hollow markers.

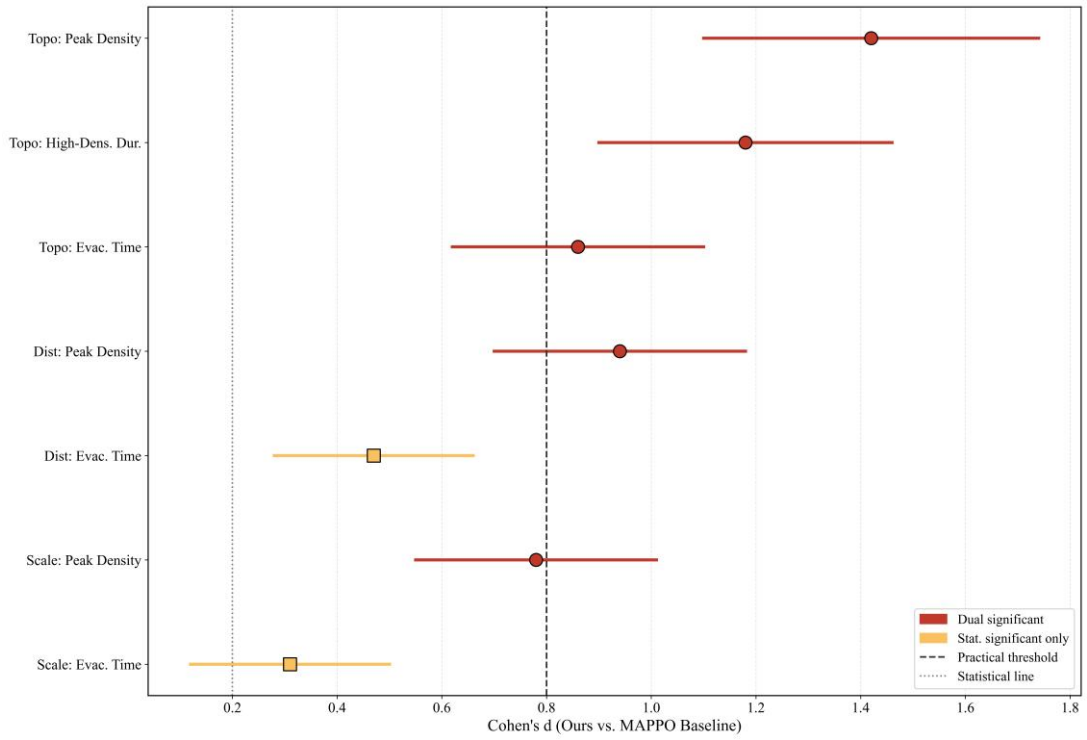


Figure 6. Forest plot showing the effect size of the proposed method compared to the MAPPO baseline

To clearly indicate which comparisons, support the claims of this paper at the statistical significance level and which merely show trends, Table 7 lists the p-values, effect sizes, and classification categories of key indicators under three perturbation dimensions. Under topological perturbation, all safety indicators and efficiency indicators achieved both statistical and practical significance; under distribution perturbation, safety indicators were significantly at both levels, while efficiency indicators were only statistically significant; under scale perturbation, peak density was significantly at both levels, while evacuation time and high-density duration showed only trends.

Table 7. Significance test results of key indicators under disturbance dimensions

Degree of disturbance	Indicators	P-value (after correction)	Cohen's d	Classification category
Topological perturbation	Peak density	<0.001	1.42	Double significance
Topological perturbation	High-density duration	<0.001	1.18	Double significance
Topological perturbation	95th percentile of evacuation time	0.004	0.86	Double significance
Distributed perturbation	Peak density	0.002	0.94	Double significance
Distributed perturbation	95th percentile of evacuation time	0.021	0.47	Only statistically significant
Scale disturbance	Peak density	0.007	0.78	Double significance
Scale disturbance	95th percentile of evacuation time	0.048	0.31	Just the trend

The effect size decreases along with the "topology-distribution-scale" direction, which is consistent with the retention rate pattern shown in Figure 4. Both of them point to the same conclusion from different statistical perspectives: The advantage of the method in this study weakens as the perturbation type shifts from structural variation to quantitative fluctuations; the p-values in the table have been corrected for multiple comparisons, and the classification is strictly based on the effect size threshold; results with only trends are not included in the claim.

4.4 Visualization of strategic behaviors and machine interpretation

The aforementioned evidence at the indicator level requires mechanism support at the behavioral level. In this section, using a typical disturbance scenario - a 30% capacity reduction in a bottleneck staircase area - as the carrier, we compare the strategic behavioral differences between the method in this paper and the MAPPO baseline. Figure 7 is a density heatmap of dual panels. The left and right panels represent the spatial density distribution of the agents of the two methods at a certain moment in the middle of the evacuation. The color shade indicates the local density. The bottleneck staircase area and its upstream passages are marked by contour lines. The differences between the two panels are concentrated upstream of the bottleneck: the density peak of the method in this paper is distributed in multiple branch channels upstream of the bottleneck, presenting a pattern of "load dispersion and moderate density at the bottleneck entrance"; while the density peak of the MAPPO baseline is concentrated in front of the bottleneck entrance, forming a typical funnel-shaped accumulation. This spatial pattern difference indicates that the diversion behavior of the method in this paper occurs before the individual reaches the bottleneck entrance, while the diversion of the baseline method can only occur after the individual is already at the bottleneck entrance and the local density has increased.

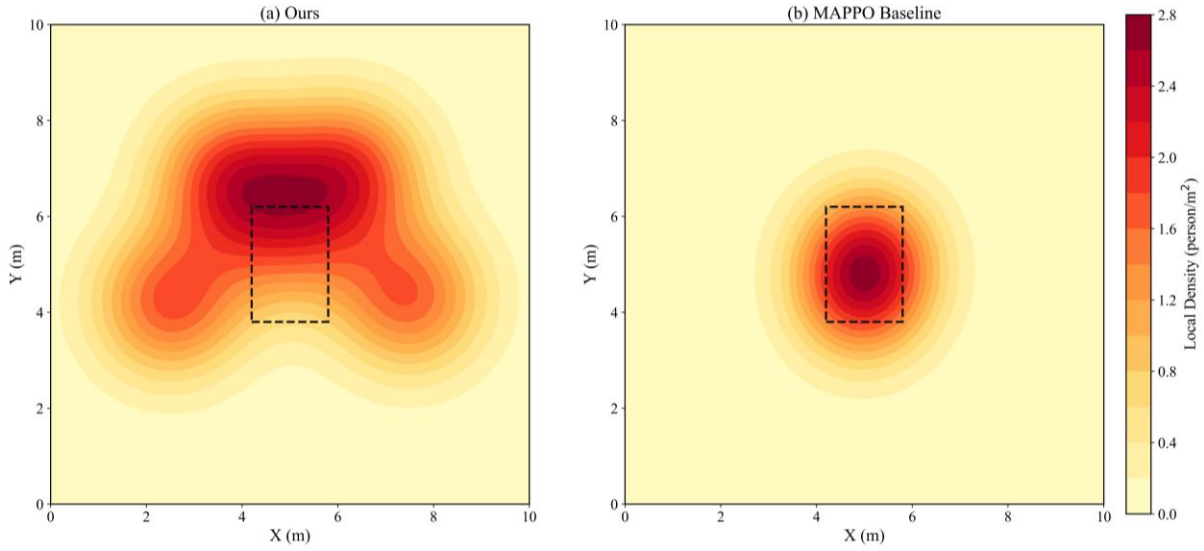


Figure 7. Comparison of agent density spatial distribution under typical disturbance scenarios

Figure 8 shows the bottleneck occupancy time curve. The x-axis represents the simulation time, and the y-axis represents the occupancy rate of the bottleneck node. The two curves represent the method in this paper and the MAPPO baseline, and the horizontal dotted line indicates the safety upper bound, while the vertical dotted line indicates the triggering time of the diversion for the two methods. The occupancy rate curve of the method in this paper shows a significant inflection point before reaching the safety upper bound and then enters a plateau phase, and its diversion triggering time is earlier than that of the baseline; the occupancy rate curve of the baseline method continues to rise and

falls back only after crossing the safety upper bound, and its diversion triggering is delayed. The time difference between the inflection points of the two curves directly measures the "predictive lead time", while the height difference of the plateau segments reflects the actual relief degree of the bottleneck pressure after diversion.

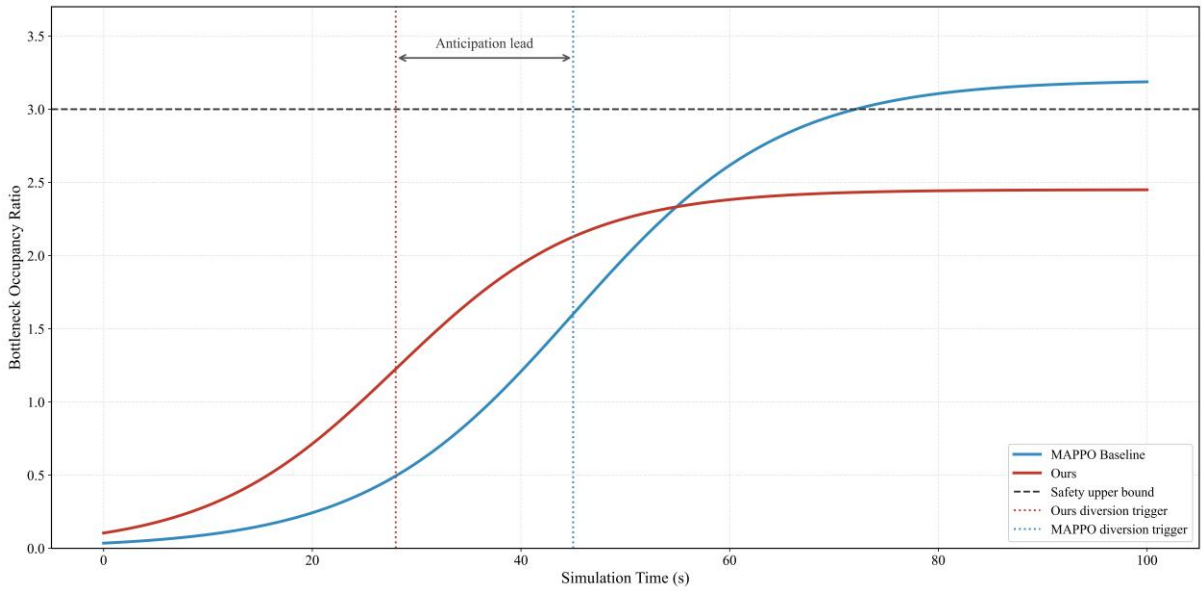


Figure 8. Comparison of bottleneck occupancy time curve and diversion trigger time

For each bottleneck node, the trigger threshold for its diversion is defined as the value at which the occupancy rate of this node first exhibits upstream diversion behavior (when the density of the upstream nodes begins to decrease and the occupancy rate of the bottleneck has not yet reached its peak) and is denoted as ρ_v^{trig}. The lower this threshold is, the earlier the strategy triggers diversion in response to the saturation of the bottleneck, and the stronger the predictability. The comparison of the diversion trigger thresholds between the proposed method and the baseline method under the same perturbation scenario is shown in Figure 9. This figure is a paired connection diagram grouped by bottleneck nodes, where each bottleneck node corresponds to a line segment connecting the thresholds of the two methods. The downward slope of the line segments indicates that the trigger threshold of the proposed method is lower. Most of the line segments slope downward and the slope is the greatest at the bottleneck nodes with the most significant pressure increase after the perturbation, indicating that the early diversion behavior of the proposed method is not uniformly applied to all bottlenecks, but is prioritized for the nodes where the pressure increases most significantly after the perturbation. This selectivity is precisely the external manifestation of the TAHSE bottleneck token attention weight at the behavioral level.

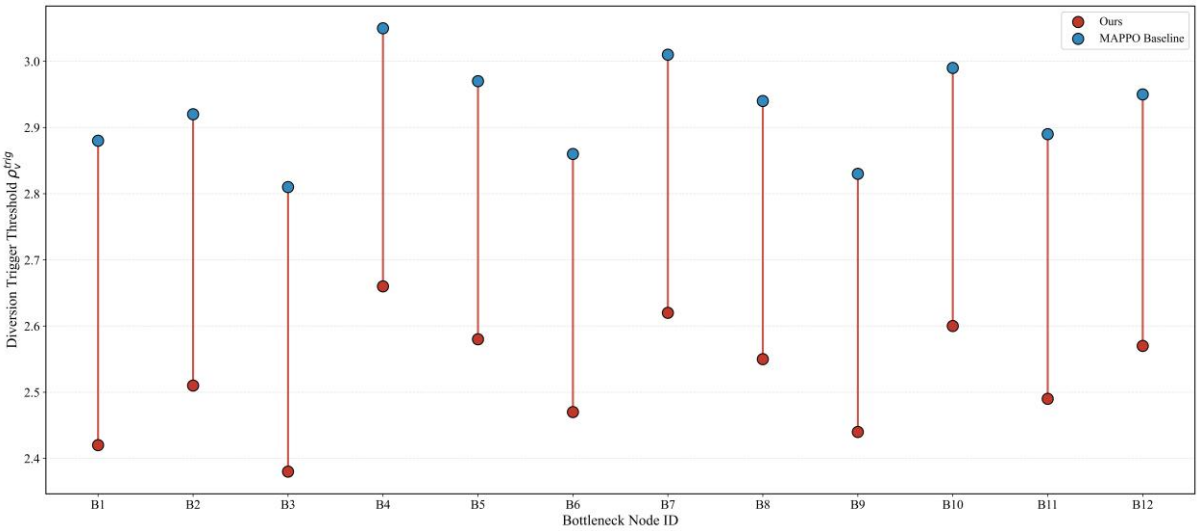


Figure 9. Comparison of diversion trigger threshold pairs at each bottleneck node

From Figures 7 to 9, the mechanism of the strategy differences can be summarized as a complete causal chain: TAHSE enables the agent to obtain the prior representation of the bottleneck situation under local observation, and BSRD enables the bottleneck pressure to receive a gradient signal at the reward level that is ahead of the safety upper limit. After alignment, the strategy learning triggers upstream diversion when the bottleneck occupancy rate exceeds the warning threshold but has not yet reached its peak, manifested as a systematic downward shift of the diversion trigger threshold; the spatial selectivity of the diversion behavior (the difference in the inclination amplitude of the nodes in Figure 9) further indicates that this behavior is driven by the attention allocation of the bottleneck token rather than a global conservative tendency.

5. DISCUSSION

The experimental evidence at the indicator level supports the combined effectiveness of TAHSE and BSRD, but this support has defined applicable boundaries that need to be clearly delineated. The bottleneck token-dependent scenario construction stage of TAHSE builds a static bottleneck graph that is fixed and its attention weights are redistributed based on the premise that the graph structure remains unchanged or the changes are known. The degradation amplitude of the method in this paper in the exit closure perturbation is less than the learning baseline, because the joint attention of bottleneck tokens and local observations can adjust "which bottleneck is more critical at the moment", rather than because the system has the ability to update the graph structure online. When the topology fails in the evacuation process and the failure information cannot be pre-injected, the misalignment between the static prior and the real structure will accumulate into a systematic bias. This limitation cannot be eliminated by increasing the types of training perturbations but requires the introduction of an online graph update mechanism. The bottleneck pressure penalty of BSRD is also constrained by the accuracy of density estimation. When there is noise in individual perception or communication limitations leading to lag in occupancy rate estimation, the triggering time of the penalty signal may deviate from the true bottleneck pressure, causing "early diversion" to degenerate into an excessive response to the lagging signal.

This work is placed within the existing lineage of MADRL evacuation research, and its positioning needs to be accurately stated. Recent studies have made substantial progress in the migration from

simulation to physical systems, using perception strategies and multi-stage curriculum training to bridge the gap between simulation and reality; another type of work handles group evacuation in complex-built environments using a knowledge-guided hierarchical architecture; and some research uses transformer observation encoders to capture the observation relationships between groups under information-limited conditions. This paper does not aim for the migration of strategies to physical robots, nor does it aim for the generalization of hierarchical control or observation encoding. Its contribution lies in transforming the bottleneck structure from the output of the simulation environment into a representational prior that the agent can carry and thereby supporting the bottleneck anticipation capability in decentralized execution. This orientation is complementary rather than competitive with the aforementioned works: if the bottleneck tokens of this paper are embedded in the underlying representation of the hierarchical architecture, or combined with the relationship modeling capability of the transformer encoder, theoretically, one can simultaneously obtain the dual advantages of structural prior and observation relationship capture, but the effectiveness of this combination still needs to be independently verified. Compared with the auditable evacuation system based on deterministic graph search, the decision-making of this paper cannot be traced back to specific sensor readings or calibration formulas, which constitutes a substantial limitation in the regulatory compliance aspect of safety-critical scenarios; however, the ability of the auditable system to respond to the dynamic evolution within the loop is precisely the design starting point of this paper's method, and the trade-off between "explainable but static" and "adaptive but partially unexplainable" between the two methods suggests the necessity of a hybrid architecture in practical deployment.

From the perspective of campus evacuation planning practice, the results of this paper provide several actionable clues. The systematic downward shift of the trigger threshold for diversion at the capacity reduction bottleneck node indicates that the learning strategy initiates upstream diversion when the bottleneck pressure has not yet exceeded the safe upper limit. The magnitude of this lead time is directly related to the attention allocation of the bottleneck token and thus can serve as a reference for evaluating "where the bottleneck warning threshold should be set if the learning strategy is adopted" during the planning stage. The behavior differentiation among seeds that occurred in the later training stage when BSRD was used alone suggests that the reward decomposition mechanism cannot work stably independently of state representation. In the absence of bottleneck perception, pressure penalties are prone to be opportunistic responses by the strategy to lower the local density reading. This diagnosis has general warning significance for any reward design that attempts to impose congestion penalties under partial observation. Moreover, the perturbation protocol of this paper shows that the method advantages decrease along the "topology - distribution - scale" direction, meaning that in scenarios dominated by fluctuations in the number of people, the marginal benefit of architectural improvement is lower than that of structural variation. The allocation of planning resources should be adjusted accordingly between these two scenarios.

The above discussion points to a more general methodological judgment: the robustness of evacuation strategies cannot be asserted by a single-dimensional perturbation test but needs to be measured separately on a perturbation set aligned with the types of scene structural variations. One of the contributions of this paper is to operationalize this judgment into a reproducible protocol, but it is also limited by the three types of perturbation dimensions selected. In real campus evacuations, the influence of crowd psychological states (panic spread, herd behavior) and information conditions (audibility of broadcast guidance, availability of mobile phone signals) on strategic behavior are not included in the perturbation design. These factors may become dominant variables in extreme events. Incorporating behavioral realism and information constraints into the perturbation dimensions while

maintaining the comparability of the statistical test framework is a necessary step to move this protocol from a methodological demonstration to a planning tool. In terms of the generalizability of the architecture, the embedding method of bottleneck tokens does not depend on the specific form of campus topology. Theoretically, it can be transferred to multi-layer heterogeneous spaces such as hospitals and transportation hubs, but the effectiveness of the transfer depends on whether the sparsity of bottleneck nodes in the target scene is comparable to that of the scenario in this paper: if the proportion of bottleneck nodes significantly increases, the computational advantage of the second-level attention will be compressed, and at this time, the information trade-offs of hierarchical encoding need to be re-evaluated.

6. CONCLUSIONS

This paper addresses two technical obstacles in emergency evacuation in universities: the spatial misalignment between local observations and global bottleneck consequences, and the inability of sparse team rewards to guide diversion behaviors. It proposes a topological-aware hierarchical state encoder and a bottleneck-sensitive reward decomposition mechanism and couples them into a unified architecture within a centralized training - decentralized execution framework. TAHSE aggregates the first-layer features using local neighborhood graph attention and uses learnable bottleneck tokens as the keys and values for the second-layer cross-domain attention, enabling agents to obtain a compressed prior on the global bottleneck distribution solely based on their own observations and pre-stored bottleneck graphs. Its computational complexity is reduced from the square level of full-graph attention to the linear level, providing a structural prerequisite for real-time reasoning in large-scale campus scenarios. BSRD decomposes a single negative time team reward into three components: progress, bottleneck pressure, and coordination completion. It triggers pressure penalties at a warning threshold lower than the safety upper limit and adjusts the safety constraints through adaptive coefficients to obtain greater weights as the scene difficulty increases. The three components enter the temporal difference objective in the value decomposition respectively, avoiding the cancellation in credit allocation. The alignment of the two designs in structure - the bottleneck tokens provide the basis for "what to see", and the pressure penalties provide the gradient signal for "how to respond" - constitutes the mechanism source of the super additive interaction effect in the ablation experiments.

The empirical part tested the above design using three-level campus scenarios and structured perturbation protocols. Under the baseline conditions, the proposed method achieved both efficiency and safety advantages simultaneously in the composite area and the entire campus scenario. The gap ratio compared to the upper bound reference was controlled within 15% in all three scenarios, while the gap ratio of the learning baseline increased from 18.4% to 31.5% as the scale expanded. The ablation experiments decomposed the total gain into the independent effects and interaction effects of two contributions: the absence of TAHSE mainly led to the degradation of safety indicators, and the absence of BSRD mainly led to the tail delay of efficiency indicators. The combination of the two showed synergistic effects that exceeded additivity in the congestion propagation range. In the perturbation test, the performance retention advantage of the proposed method under topological perturbation was the most prominent. The effect size decreased along the "topology - distribution - scale" direction, and the systematic shift of the diversion trigger threshold on the capacity-reduced bottleneck nodes provided direct behavioral evidence for the bottleneck prediction ability. These results collectively support a limited but testable conclusion: In a multi-layer heterogeneous campus space with a basically stable bottleneck graph structure and sparse bottleneck nodes, embedding the bottleneck as a carriable prior

into the decentralized strategy and guiding the diversion with a reward signal ahead of the upper safety bound can achieve joint improvement of efficiency and safety at a controllable computational cost, and this improvement is more reliable under structural variations than under quantitative fluctuations.

The limitations of the study and its subsequent directions are also clearly defined. The static fixation of the bottleneck graph limits the method's adaptability in scenarios with unknown topological structures. Online graph updates and the introduction of dynamic bottleneck tokens are direct paths to eliminating this limitation. The protocol has not yet covered the dimensions of human psychology and information conditions. Incorporating behavioral realism into structured perturbations while maintaining the comparability of the statistical framework is a key step in moving the evaluation protocol from a methodological demonstration to a planning tool. The feasibility of multi-campus transfer learning depends on the sparsity of bottlenecks in the target scenario and the matching degree with the source scenario. When the matching degree is insufficient, the migration of bottleneck tokens may require the redesign of hierarchical encoding information selection. Additionally, the decisions made by the method cannot be traced back to specific environmental readings, which constitutes an inherent disadvantage compared to deterministic systems in security-critical scenarios. The hybrid architecture-generating diversion tendencies through learning strategies and implementing traceable path allocation through deterministic rules-may be a pragmatic direction that balances adaptability and auditability.

Abbreviations

TAHSE, Topological Awareness Hierarchical State Encoder;

BSRD, Bottleneck Sensitive Reward Decomposition Mechanism;

CTDE, Centralized Training–Decentralized Execution.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **L.Y.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Funding information

This research was supported by the Youth Science and Technology Innovation Project of Lanzhou Institute of Technology (No. 2024KJ-32).

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **L.Y.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used DeepSeek for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Zhang, M., Li, Z., Si, H., Cheng, L., Zhou, X., & Wang, B. (2023). Urban travel time and residential location choice: The impacts of traffic congestion. Sustainable Cities and Society, 99, 104975. https://doi.org/10.1016/j.scs.2023.104975

[2] Fan, M., Hedayati Marzbali, M., Abdullah, A., & Maghsoodi Tilaki, M. J. (2024). Using a space

syntax approach to enhance pedestrians' accessibility and safety in the historic city of George Town, Penang. Urban Science, 8(1), 6. https://doi.org/10.3390/urbansci8010006

[3] Chen, L., Deng, P., Li, L., & Hu, X. (2025). Mixed motivation driven social multi-agent reinforcement learning for autonomous driving. IEEE/CAA Journal of Automatica Sinica, 12(6), 1272-1282. https://doi.org/10.1109/jas.2025.125201

[4] Lokam, A., & Dasari, K. (2026). A QMIX-driven routing protocol for dynamic Underwater Acoustic Sensor Networks. Scientific Reports. https://doi.org/10.1038/s41598-026-74771-2

[5] Guan, Y., Ren, Y., Sun, Q., Li, S. E., Ma, H., Duan, J., ... & Cheng, B. (2022). Integrated decision and control: Toward interpretable and computationally efficient driving intelligence. IEEE transactions on cybernetics, 53(2), 859-873. https://doi.org/10.1109/tcyb.2022.3163816

[6] Malathi, V., Sreedharan, P., PR, R., Anil Kumar, V., Sadasivan, A. L., Udupa, G., ... & Troppina, A. (2025). Decision-making for path planning of mobile robots under uncertainty: A review of belief-space planning simplifications. Robotics, 14(9), 127. https://doi.org/10.3390/robotics14090127

[7] Guo, Y., Mao, Z., Chen, X., Fan, X., Han, Y., Su, X., ... & Tian, H. (2026). Floor field cellular automata model of pedestrian crush and stampede behavior based on crowding force transmission mechanism. Physics Letters A, 131951. https://doi.org/10.1016/j.physleta.2026.131951

[8] Zambare, P., & Liu, Y. (2025). Modeling and simulation of pedestrian behavior in traffic environment: A systematic review of approaches and software. IEEE Access. https://doi.org/10.1109/access.2025.3602471

[9] Silva, E. C., Damazo, G. S., & Martins, L. G. (2025, February). Cellular Automata Applied to the Simulation of Pedestrian Evacuation with Surmountable Obstacles and Route Alternation Through an Impatience Mechanism. In International Conference on Agents and Artificial Intelligence (pp. 562-582). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-25032-2_27

[10] Imtiaz, T., Chai, L., Heal, K., Luo, X., Park, J., Dy, J., & Flynn, J. (2025, December). LVT: Large-Scale Scene Reconstruction via Local View Transformers. In Proceedings of the SIGGRAPH Asia 2025 Conference Papers (pp. 1-12). https://doi.org/10.1145/3757377.3763838

[11] Sun, T., Ji, J., Li, J., & Xv, Z. (2026). SpectVit: Spectral-spatial gating for efficient vision transformers. Pattern Recognition, 114749. https://doi.org/10.1016/j.patcog.2026.114749

[12] Ge, L., Wang, S., & Wang, G. (2024). Rapid vehicle trajectory prediction based on multi-attention mechanism for fusing multimodal information. Electronics, 13(23), 4806. https://doi.org/10.3390/electronics13234806

[13] Han, D., Ye, T., Han, Y., Xia, Z., Pan, S., Wan, P., ... & Huang, G. (2024, September). Agent attention: On the integration of softmax and linear attention. In European conference on computer vision (pp. 124-140). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72973-7_8

[14] Ye, K., Wu, Z., Wu, Y., Qian, H., Woo, S., Srigrarom, S., & Khoo, B. C. (2026). Topology-Safe Air-to-Air Interception via "Go"-Inspired Metrics and Multi-Agent Reinforcement Learning. Aerospace, 13(8), 680. https://doi.org/10.3390/aerospace13080680

[15] Terry, J. K., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., ... & Ravi, P. (2021). Pettingzoo: Gym for multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 15032-15043. https://doi.org/10.48550/arXiv.2009.14471

[16] Metelo, F., Soares, C., Racković, S., & Costa, P. Á. (2024, August). Peersimgym: An environment for solving the task offloading problem with reinforcement learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 38-54). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70378-2_3

[17] Yang, H., Shi, J., Jiang, N., Yu, H., & Yang, L. (2026). Congestion-Mitigating Evacuation Path Planning in Subway Stations via Graph-Theoretic Deep Reinforcement Learning. Simulation Modelling Practice and Theory, 103315. https://doi.org/10.1016/j.simpat.2026.103315

[18] Moreno, J. V., Gonçalves, A. B., & Falcão, A. P. (2026). Bridging BIM and Path-Finding Algorithms: A Systematic Review and Analysis of Escape Route Generation. Sustainability, 18(20), 10490. https://doi.org/10.3390/su182010490

[19] Zhong, Z., & Ren, Z. (2026). Multi-agent Reinforcement Learning Optimization for Urban Earthquake Emergency Evacuation with GIS-based Real-time Decision Support. Journal of Earthquake and Tsunami, 2540002. https://doi.org/10.1142/s1793431125400020

[20] Low, D., & Zhou, Y. (2025). Cooperative multi-agent reinforcement learning for robotic systems: A review. Multiagent and Grid Systems, 21(2), 96-123. https://doi.org/10.1177/15741702251370050

[21] Zhang, J., & Wu, S. (2026). Master-Refined MAPPO for Long-Term Joint Resource Scheduling in NOMA-MEC Systems. Symmetry, 18(7), 1243. https://doi.org/10.3390/sym18071243

[22] Torres Rodríguez, I. J., Marsà Fargas, J., Candelo Zuluaga, C., Pérez Díaz, C. A., Martin Hernández, L. D., Sanz Merodio, D., ... & López Estévez, M. (2026, May). Toward Reproducible Actuator Characterization: A Comprehensive Framework for Robotics Applications. In Actuators (Vol. 15, No. 6, p. 301). MDPI. https://doi.org/10.3390/act15060301

[23] Kalkanlı, D., Kundak, S., Atun, F., & Van Westen, C. J. (2026). Human-Centric Digital Twins for Spatial Sustainability: A Procedural VR Framework for Calibrating Agent-Based Evacuation Models in Diverse Urban Morphologies. Sustainability, 18(3), 1482. https://doi.org/10.3390/su18031482

APA Citation

Yang, L. (2026). Research on Simulation and Path Optimization of University Emergency Evacuation Based on Multi-Agent Reinforcement Learning. *Journal of Digital Frontier, 1*(2), 124-152. **https://doi.org/10.67541/jdf2612**