Dear Editors and Reviewers,

Thank you for your insightful comments and constructive feedback on our manuscript. We have carefully considered all your suggestions and have revised the manuscript accordingly. We believe these changes have significantly improved the clarity, rigor, and overall quality of our paper.

Please find our point-by-point responses to your comments below.

Comment 1: *Please explain how you determine the transition probabilities. Did authors use any real data to train the transition probabilities value?*

Response to Comment 1:
Thank you for this important question. We have added a detailed clarification **in Section 3.3** of the revised manuscript.

In our study, the transition probability is not trained from data but is determined by the nature of the environment model. It is important to clarify the nature of the transition probability in our model. In a general Markov Decision Process (MDP), the state transition function can be stochastic, representing inherent uncertainty in an environment's dynamics. In contrast, **the raster-based environment in this study is deterministic**. Specifically, any action taken from a given state (a grid cell) deterministically leads to a single, known subsequent state (the adjacent cell). This means the transition probability is 1 for the resulting state and 0 for all others. As the environment's transition model is perfectly known and deterministic, **it does not need to be learned or estimated from real data.** The agent's learning challenge is therefore focused not on modeling the environment's dynamics, but on discovering the optimal policy within these known dynamics.

Comment 2: *Please find the reference to the reward function structure. Where to the values in Table 4 come from? Please give a detailed explanations.*

Response to Comment 2:
Thank you for the valuable suggestion. We have significantly expanded the explanation of our reward function design in Section 3.4 and included relevant citations.

The reward function detailed in Table 4 was designed heuristically based on the principle of **reward shaping** (Ng et al., 1999; Ibrahim et al., 2024), a common approach for creating dense

and informative signals to accelerate learning. **The specific values were calibrated empirically against the environment's scale.** Our analysis indicated that a medium-length evacuation route comprises approximately 100 steps, which established the step reward as a baseline unit. The destination reward (+100) was thus set to counteract the cumulative penalty of a medium-length path, while risk penalties were scaled significantly higher to prioritize safety over minor efficiency gains. This empirically-grounded calibration proved highly effective in guiding the agent toward optimal and safe evacuation routes.

Comment 3: *Please add a citation to Figure 7. It is a very generous RL model figure, so you need to cite the references.*

Response to Comment 3:
Thank you for pointing this out. Figure 7 indeed illustrates the standard reinforcement learning framework. We have revised the caption for Figure 7 to properly credit the foundational work in the field. The caption now begins with: "The DRL model, **adapted from Sutton and Barto**."

Comment 4: *Please explain how your deep learning network is trained. Please specify the input data and output results. The results should validate why deep learning is imperative for this research.*

Response to Comment 4:
Thank you for this insightful comment, which has prompted us to significantly improve the structure and clarity of our methodology section. We have reorganized Section 3.4 to first explain why a deep learning approach is imperative, and then detail how the network is trained.

1. Justification for Deep Learning: We have added a new paragraph at the beginning of Section 3.4 to address this. Deep reinforcement learning is imperative for this research for several key reasons.

Firstly, the problem is characterized by a **high-dimensional state space**. The agent's state is not a simple coordinate but a rich, image-like observation, rendering traditional tabular methods computationally infeasible. More importantly, a deep neural network provides powerful **generalization**, learning abstract environmental features—such as intersections or flooded zones—rather than memorizing individual states. This allows the agent to make intelligent decisions in novel, previously unseen situations. Lastly, its end-to-end learning capability allows for policy optimization directly from raw environmental data, bypassing manual feature engineering.

2. Input, Output, and Training Process: We have also added a new, consolidated paragraph that clearly defines the training mechanics.

Input: At each step, the network takes the agent's local observation as input—a multi-channel tensor of size (2rob+1,2rob+1,4) containing information on roads, shelters, risks, and distances. **Output:** The output is a vector of 8 Q-values, predicting the expected return for each possible action. **Training:** The network's weights are optimized by minimizing the Mean Squared Error between the predicted Q-values and target Q-values. To ensure stable training, this process utilizes a separate target network and samples experiences from a replay memory to break temporal correlations.

We believe this new structure better addresses your questions and improves the overall readability of the manuscript.

Once again, we sincerely thank you for your time and for providing such valuable feedback to help us improve our work.

Sincerely,

Chuanfeng Liu, Yan Li, Hao Qin, Wenjuan Li, Lin Mu, Si Wang, Darong Liu, and Kai Zhou