# 拡散モデルを利用した深層強化学習による歩行者混雑環境下での移動ロボットナビゲーション

第 3 報　アニーリングによる拡散モデルの性能向上

○富田 湧（九州大学），松本 耕平（九州大学），兵頭 侑樹（九州大学），
長久陽斗（九州大学），倉爪 亮（九州大学）

# Mobile Robot Navigation in Crowded Environments via Diffusion-Based Reinforcement Learning

○ Yuki Tomita (Kyushu Univ.), Kohei Matsumoto (Kyushu Univ.), Yuki Hyodo(Kyushu Univ.),
Haruto Nagahisa (Kyushu Univ.) and Ryo Kurazume (Kyushu Univ.)

Abstract: In recent years, mobile robot navigation has been applied to various real-world scenarios, where safe and efficient navigation in dynamic and crowded environments is essential. Deep reinforcement learning (DRL) has been widely studied for this purpose, but conventional policy models often rely on unimodal Gaussian distributions, which limit behavioral diversity and lead to premature convergence in complex tasks. To address this issue, we propose using a diffusion model as a policy representation in DRL, enabling flexible and multimodal action generation. In addition, we introduce an annealing-based temperature control mechanism that gradually adjusts the inverse temperature parameter during training. This approach encourages exploration in the early stage and progressively lowers the temperature in later stages to converge the distribution. Through this process, the robot can leverage multimodal exploration while ultimately acquiring a high-performance policy. Experiments in both simulation and real-world robot navigation confirm that the proposed method improves learning efficiency, utilizes diverse action patterns, and achieves robust performance in crowded environments.

## 1.　緒言

近年，人手不足を背景に様々な分野で移動ロボットの導入が進められている．特に病院やショッピングモールなど人が多く存在する環境への導入においては，歩行者が行き交う混雑状況でも衝突を回避しながらスムーズに移動できる手法が不可欠である．このような歩行者混雑環境下におけるロボットナビゲーションはソーシャルナビゲーションと呼ばれ，深層強化学習を利用した手法が数多く提案されている[1)~3)]．深層強化学習は，シミュレーションによって環境を再現することで人的リソースを消費せずに繰り返し学習できる．さらに，得られた方策を用いてデータを収集し，逐次的に性能を改善することで，歩行者の不規則な行動に対しても経験的に多様な状況へ適応可能であり，有効なアプローチである．しかし，従来手法の多くは行動生成を行う方策に単峰的な正規分布を用いており，その結果，生成される行動は偏りが生じやすい．この制約により，歩行者混雑環境下のように複数の局所最適解が存在する複雑なタスクでは多様な行動パターンを十分に網羅することができず，新たな経験の獲得が困難となり，学習が早期に停滞する可能性が指摘されている[4)]．

一方，近年，画像生成などの分野で注目を集めている拡散モデルは，段階的にデータにノイズを付与し，これを除去する逆拡散過程を学習することで，多様で高品質なデータを生成可能である[5)]．先行研究では，この特性を方策表現に応用することで，従来の単峰的な分布に比べて複雑な行動モードを柔軟に表現できることが確認されている[6)]．さらに，事前にコスト関数を設計し，学習後のモデルに設計したコスト関数に基づく条件付けを行い，学習時とは異なる静的障害物を含む状況環境においても適応可能であることが報告されている[7)]．

$$\pi_\theta(\boldsymbol{s}_t, \boldsymbol{a}_t) = \frac{\beta Q_\phi(\boldsymbol{s}_t, \boldsymbol{a}_t)}{\int_{\boldsymbol{a}_t \in A} \beta Q_\phi(\boldsymbol{s}_t, \boldsymbol{a}_t)}$$

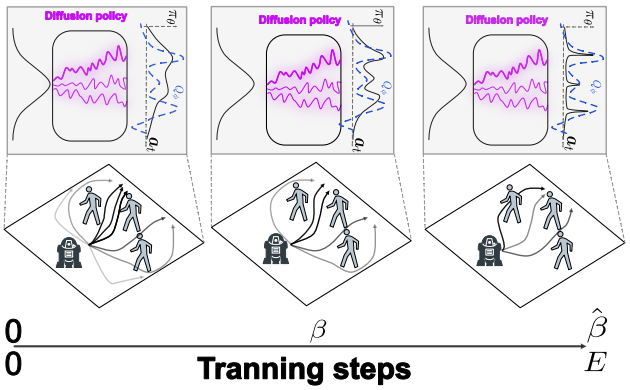


**Fig. 1**　Conceptual diagram of the proposed method. The diffusion-based policy generates diverse actions, with $\beta$ annealed during training to promote exploration in the early step and converge the distribution in later steps. The maintained multimodality further enables guidance-based conditional action generation for adaptive navigation in crowded environments.

そこで本研究（第 3 報）では，拡散モデルを強化学習における方策モデルとして利用する．さらに，Fig. 1 に示すように，学習の進行に応じて温度パラメータを段階的に制御する手法を導入する．この制御には二つの目的がある．一つ目は，初期段階において探索を促進し学習性能を向上させることである．二つ目は，学習初期に分布が一箇所に偏ってしまうことで起きる

モード崩壊を抑制し，拡散モデルが本来持つ多峰性を保持することで，ガイダンスを用いた条件付けによる行動生成の性能を高めることである．これにより，多様な行動を保持しつつ方策を最適化することを可能とし，歩行者混雑環境下におけるロボットナビゲーションの性能向上を目指す．

本研究の貢献は以下のとおりである．

- 逆温度パラメータを段階的に制御するアニーリング手法を導入することで，性能の向上が確認した．
- アニーリングによってモード崩壊を抑制し，多峰性を保持することで，ガイダンスを用いた条件付けによる行動生成の性能を向上できることを確認した．
- 提案手法を実機ロボットに適用することで，シミュレーション環境にとどまらず実環境においても有効性を確認した．

## 2. 拡散モデルを用いたナビゲーション

### 2.1 拡散モデル

拡散モデルは, 拡散過程と逆拡散過程から構成される．拡散過程では目標分布から得られるサンプル $\boldsymbol{x}^0$ に少しずつノイズを付与していくことで, 最終的にサンプルを完全なノイズにする．逆拡散過程では，ノイズの付与されたサンプル $\boldsymbol{x}^\tau$ から学習モデルを用いてノイズ除去を行う．この 2 つの過程をタイムステップ $\tau$ を変更しながら反復的に行うことで完全なノイズから目標サンプルを生成する．

拡散モデルの生成過程, 逆拡散過程を式 (1),(2) に示す．ここで，$\alpha_\tau$ はサンプルにノイズを付与する割合を決定するハイパーパラメータである．

$$\boldsymbol{x}^\tau = \sqrt{\alpha_\tau}\boldsymbol{x}^0 + \sqrt{1-\alpha_\tau}\boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{1}$$

$$\boldsymbol{x}^0 = \frac{\boldsymbol{x}^\tau - \sqrt{1-\alpha_\tau}\boldsymbol{\epsilon}_\theta}{\sqrt{\alpha_\tau}}. \tag{2}$$

ここで,$\boldsymbol{\epsilon}_\theta$ は学習されるモデルの出力である．学習は，式 (3) のように，ノイズの付与されたサンプルから付与されたノイズを推定し，二乗誤差を最小化する．

$$L(\theta) = \mathbb{E}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^\tau, \tau)||^2]. \tag{3}$$

この学習過程を通してノイズを予測し，式 (2) で段階的にノイズ除去を行うことで，完全なノイズから多様なデータを確率的に生成することができる．

### 2.2 拡散モデルを用いた深層強化学習

本研究では，条件付き拡散モデルを方策として利用し，深層強化学習を行う．学習アルゴリズムには，拡散モデルに Actor-Critic のフレームワークを利用した Q スコアマッチング[8] を採用し，行動価値関数 $Q_\phi(\boldsymbol{s}_t, \boldsymbol{a}_t)$ と方策 $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)$ を同時に学習する．行動価値関数の学習においては式 (4) のようにベルマン方程式によって得られる誤差を最小化する．ここで, $r_t(\boldsymbol{s}_t, \boldsymbol{a}_t)$ は時刻 $t$ における報酬, $\gamma$ は割引率,$\boldsymbol{s}_{t+1}$ は次の状態を表す．

$$L(\phi) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim D}\Big[\big(Q_\phi(s_t,a_t) - (r(s_t,a_t) + \gamma \max_{a'} Q_{\phi'}(s_{t+1},a'))\big)^2\Big]. \tag{4}$$

このように，エピソード内の時刻をもとに行動価値関数を再帰的に学習することで，エピソード開始地点から特定の時刻 $t$ までに得られる報酬の期待値を推定することができる．方策の更新においては方策分布が式 (5) のボルツマン分布に従うように拡散モデルにおけるスコアマッチングの更新式 (6) を変形し，

$$\pi_\theta(\boldsymbol{s}_t, \boldsymbol{a}_t) = \frac{\beta Q_\phi(\boldsymbol{s}_t, \boldsymbol{a}_t)}{\int_{a\in A} \beta Q_\phi(\boldsymbol{s}_t, \boldsymbol{a}_t)} \tag{5}$$

$$L_{\mathrm{SM}} = \mathbb{E}[\|\Psi_\theta(\boldsymbol{s}_t, \boldsymbol{a}_t, \tau) - \nabla_{\boldsymbol{a}} \log \pi_\theta(\boldsymbol{s}_t, \boldsymbol{a}_t)\|^2]. \tag{6}$$

式 (7) を最小化するように学習する．ここで，学習対象となる $\Psi_\theta$ は確率密度関数の勾配であるスコアを近似的に学習し，そのスコアに基づくノイズ除去を繰り返すことで行動を生成する．これにより，行動価値関数が高い行動を生成するように拡散モデルを更新する．ここで，$\beta$ は式 (5) における逆温度パラメータであり，ボルツマン分布が行動価値関数に対してどの程度鋭敏に応答するかを制御する．逆温度パラメータ $\beta$ を小さく設定すると行動のランダム性が高まり，多様な探索が促進される．一方で大きく設定すると高価値な行動に分布が集中し，探索よりも最適行動の選択が優先される．

$$L_{\mathrm{QSM}} = \mathbb{E}[\|\Psi_\theta(\boldsymbol{s}_t, \boldsymbol{a}_t, \tau) - \beta\nabla_{\boldsymbol{a}_t} Q_\phi(\boldsymbol{s}_t, \boldsymbol{a}_t)\|^2]. \tag{7}$$

このように, 拡散モデルを学習することで, 行動価値に基づく複雑な確率分布を柔軟に表現でき, 多様な行動を生成可能な方策を学習できる．

### 2.3 ガイダンスを用いた拡散モデルの行動生成

拡散モデルは，ベイズ則に基づく条件付きスコア関数を利用することで，学習後に新たな条件に従うよう行動生成を誘導することができる．

式 (8) に, ベイズ則に基づいた拡散モデルの条件付きスコア関数を示す．

$$\nabla_{\boldsymbol{x}^\tau} \log p(y|\boldsymbol{x}^\tau, \boldsymbol{s}_t) = \nabla_{\boldsymbol{x}^\tau} \log p(\boldsymbol{x}^\tau, \boldsymbol{s}_t) + \lambda\nabla_{\boldsymbol{x}^\tau} \log p(\boldsymbol{x}^\tau|y, \boldsymbol{s}_t). \tag{8}$$

ここで，$\lambda > 0$ はガイダンススケールと呼ばれ，値が大きいほど条件 $y$ の影響を強める．第一項は学習済み拡散モデルのスコアで計算でき，第二項は Clean Guidance[9] に基づき事後分布を近似し，エネルギー関数 $J(\hat{\boldsymbol{x}}_0)$ を用いて壁回避などの制約を導入することで，式 (9) のように定式化される．

$$\nabla_{\boldsymbol{x}^\tau} \log p(\boldsymbol{x}^\tau, \boldsymbol{s}_t) = \Psi_\theta(\boldsymbol{s}_t, \boldsymbol{x}^\tau, \tau). \tag{9}$$

$$\lambda\nabla_{\boldsymbol{x}^\tau} \log p(\boldsymbol{x}^\tau|y, \boldsymbol{s}_t) = \lambda\nabla_{\boldsymbol{x}^\tau} J(\hat{\boldsymbol{x}}^0). \tag{10}$$

先行研究では，$J(\hat{\boldsymbol{x}}^0)$ を式 (11) のように定義している．ここで，$\boldsymbol{p}_t^w$ はロボット移動後の壁の最近傍点，$\boldsymbol{p}_t^r$ はロボット位置を表し，$\hat{\boldsymbol{x}}^0$ は各生成ステップにおいて推定されるノイズ除去後の行動である．この行動 に従ってロボットが移動した際の位置と壁との距離に基づきコストを算出し，これを用いて壁から離れるように条件付けを行う．

さらに，適切にガイダンススケール $\lambda$ を適切に設定

することで，壁と歩行者の両方を回避しつつ，効率的にゴールへ到達するナビゲーションを実現できることが示されている．Fig. 2 に，拡散モデルの行動生成にガイダンスを用いて条件付けを行い，静的障害物を回避する手法の概念図を示す．

$$J(\hat{\boldsymbol{x}}^0) = \exp\left(-\|\boldsymbol{p}_t^w - (\boldsymbol{p}_t^r + \Delta t \hat{\boldsymbol{x}}^0)\|_2\right). \tag{11}$$

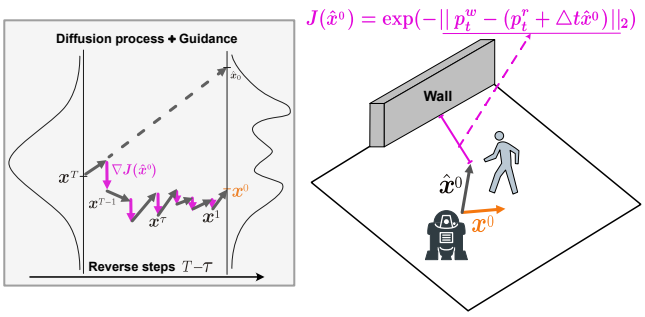


**Fig. 2** Image of diffusion model and action generation using guidance

ここで，ガイダンスにおいては，モデルが有する方策分布の多峰性が高いほど，誘導可能な行動の選択肢が増加する．その結果，生成過程を誘導した際に，歩行者に衝突することなく壁を回避し，ゴールへ到達できる．

## 3. 提案手法

### 3.1 問題設定

本研究では，X-Y 平面の歩行者が複数存在する環境で，ロボットが衝突を回避しながら目的地へ到達するタスクを考える．報酬，状態，行動を以下のように定義する．

- 状態：歩行者とロボットの位置と速度データを状態として扱う．各歩行者，ロボットの観測はベクトル $(p_x^i, p_y^i, v_x^i, v_y^i)$ であり，$i$ 番目の歩行者またはロボットにおいて，$(p_x^i, p_y^i)$ は位置，$(v_x^i, v_y^i)$ は速度を表す．
- 行動：ホロノミックな全方位移動ロボットを想定し，2 次元空間におけるロボットの x 軸方向の入力速度 $v_x$ と y 軸方向の入力速度 $v_y$ からなる 2 次元ベクトル $(v_x, v_y)$ を用いる．
- 報酬：式 (12) に本研究における報酬を示す．$d_t$ はロボットと周囲の歩行者間の最小距離を表し，$\boldsymbol{p}_t^r$ は時刻 $t$ におけるロボットの位置，$\boldsymbol{p}_g$ はロボットの目標位置を示す．

$$r(\boldsymbol{s}_t, \boldsymbol{a}_t) = \begin{cases} -0.25 & \text{if } d_t < 0 \\ -0.1 + d_t/2 & \text{else if } d_t < 0.2 \\ 1 & \text{else if } \boldsymbol{p}_t^r = \boldsymbol{p}_g \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

式 (12) では，設計者のバイアスを避けるためゴールとの距離は用いず，衝突，ゴール到達，歩行者近傍の通過時のみ報酬が変化する疎な報酬関数を採用する．

### 3.2 学習手法

本研究で用いる学習アルゴリズムを Algorithm. 1 に示す．従来手法と異なり，本手法では逆温度パラメータ $\beta$ をアルゴリズム内で示したアニーリングスケジューラ $f(n; \hat{\beta}, E)$ に基づいて逐次更新する．スケジューラとしては線形，指数関数，シグモイド関数などを選択でき，いずれも $\beta = 1$ から目標値 $\beta = \hat{\beta}$ に向かって変化するよう設計されている．アニーリングスケジューラを Table.1 に示す．

**Table 1** Definitions of annealing schedulers

| Scheduler | Function $f(n; \hat{\beta}, E)$ |
|---|---|
| Linear | $\hat{\beta} \cdot \frac{n}{E}$ |
| Linear + Clip ($\eta < 1$) | $\min\left(\hat{\beta} \cdot \frac{n}{\eta E}, \hat{\beta}\right)$ |
| Exponential | $\hat{\beta}^{\frac{n}{E}}$ |
| Sigmoid | $\hat{\beta} \cdot \frac{\hat{\beta} - 1}{1 + e^{-k\frac{n - E/2}{E}}}$ |

これにより，学習初期には探索を促進して学習効率を向上させ，学習が進むにつれて行動価値関数 $Q(\boldsymbol{s}_t, \boldsymbol{a}_t)$ に対する反応を強めることで，方策が獲得する分布のモード崩壊を抑制しつつ多峰性を保持することが可能となる．さらに，多峰性を維持することで，学習後のガイダンスを用いた条件付けによる行動生成の性能を高めることができる．本研究ではこれらのスケジューラの中で歩行者人数の変化に対して，性能が最も高いモデルを汎化性の高いモデルとして使用する．

**Algorithm 1:** Training algorithm using QSM

Initialize score network $\Psi_\theta$ and critic networks $Q_{\phi^1}$ and $Q_{\phi^2}$
Set parameter values of the target critics $Q_{\phi'^1}$ and $Q_{\phi'^2}$ equal to those of the main critics
**for** $n = 1$ **to** $E$ **do**
  Explore using the policy until finishing an episode
  After finishing an episode, store the trajectory of $(\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}'_t)$ to the replay buffer $\mathcal{D}$
  Sample a batch $\mathcal{B} = (\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}'_t)\}$ from the replay buffer $\mathcal{D}$
  Sample actions for computing targets $\hat{\boldsymbol{a}}'_t \sim \pi_\theta(\cdot \mid \boldsymbol{s}'_t)$
  Calculate targets for the Q-function $y(r_t, \boldsymbol{s}'_t) = r_t + \gamma\left(\min_{i=1,2} Q_{\phi'^i}(\boldsymbol{s}'_t, \hat{\boldsymbol{a}}'_t)\right)$
  Update critics by minimizing $L_{\text{critic}} = \frac{1}{|\mathcal{B}|}\sum(Q_{\phi^i}(\boldsymbol{s}_t, \boldsymbol{a}_t) - y(r_t, \boldsymbol{s}'_t))^2 \quad \text{for } i = 1, 2$
  Create noisy action by (1) $\boldsymbol{x}^\tau = \sqrt{\alpha_\tau}\boldsymbol{a}_t + \sqrt{1-\alpha_\tau}\boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
  Update score network by minimizing $L_{\text{QSM}} = \frac{1}{|\mathcal{B}|}\sum\|\Psi_\theta(\boldsymbol{s}_t, \boldsymbol{x}^\tau, \tau) - \beta\nabla_{\boldsymbol{x}^\tau} Q(\boldsymbol{s}_t, \boldsymbol{x}^\tau)\|^2$
  Update the target critic by polyak averaging $\phi'^{,i} \leftarrow \rho\phi'^{,i} + (1-\rho)\phi_i \quad \text{for } i = 1, 2$
  Update $\beta$ by annealing scheduler
  $\beta \leftarrow f(n; \hat{\beta}, E)$
**end**

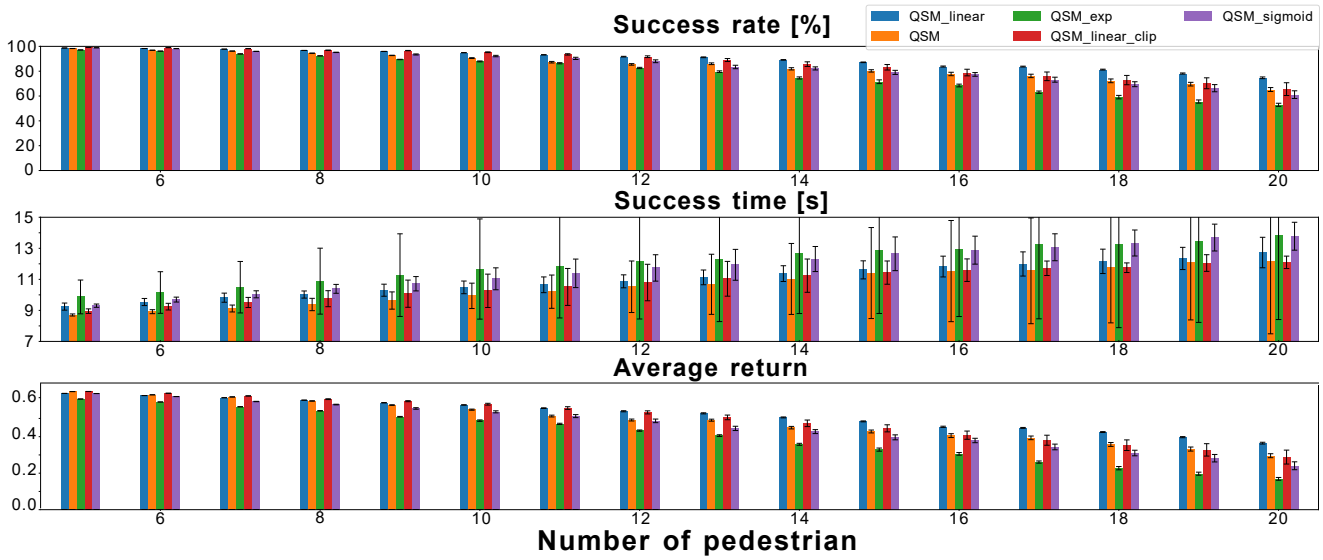


**Fig. 3** Comparison of annealing schedulers across pedestrian numbers.

## 4. シミュレーション実験

### 4.1 シミュレーション環境

シミュレーション実験では，CrowdNav 環境の circle crossing シナリオを利用する[10]．このシナリオでは，ロボットは初期位置 $(x, y) = (0, -4)$ からゴール地点 $(x, y) = (0, 4)$ を目指して進む．歩行者は ORCA[11] に従って行動し，初期化時に半径 4 m の円上にランダムに配置される．学習はロボットを ORCA に基づいて動作させ，2,000 エピソード分の遷移情報を事前に収集しておき，100,000 エピソード分のオンライン学習を行った．性能評価は 5 つの異なるシードで学習を行ったモデルを，歩行者人数を 5 から 20 人に変化させたシナリオで，500 パターンのテストケースを用いて評価した．ここで，逆温度パラメータとなる $\beta$ を事前に 1 から 1,000 の範囲を 100 ごとに調査した結果，最も平均収益の高かった $\hat{\beta} = 400$ をアニーリングの最終値として，実験を行う．

### 4.2 歩行者人数に対する性能評価

Fig. 3 に歩行者人数を 5 人から 20 人に変更したときの，各アニーリングスケジューラの結果を示す．結果が同じ人数の場合どの手法も性能に差はなく平均収益はパラメータを固定している QSM が最も高い．しかし，歩行者人数が増加していくと逆温度パラメータに線形アニーリングを施した QSM_linear が最も成功率，平均収益が高くなった．これにより，アニーリングを用いて学習段階に応じて分布の逆温度パラメータを変化させることで歩行者人数が多い環境では，性能が向上することを確認した．

### 4.3 アニーリングにおける多峰性評価

モデルの行動分布の多峰性を定量的に評価するため，本研究では前節で最も性能の高かった QSM および QSM_linear から，同一の状態において 1,000 回の行動サンプルを取得し，その分布を解析した．1 次元の行動成分については，カーネル密度推定（KDE）を用い，ピークの数を数えることでモード数を推定した[12]．2 次元の行動全体については，サンプル行動に対してガウス混合モデル（GMM）を適用し，ベイズ情報量規準（BIC）を最小化する成分数を多峰性の指標として採用した[13]．各ステップの多峰性スコアはエピソード内の全ステップで平均化し，エピソード単位のスコアを算出した．さらに複数のエピソードに対して平均および標準偏差を計算することで，シナリオ全体における各ステップの方策の多峰性の傾向を定量的に評価した．Table. 2 に QSM と QSM_linear の circle crossing シナリオにおける多峰性を比較した結果を示す．

**Table 2** Evaluation of multimodality in distributions.

| Method | $all$ | $all^{2D}$ | $x$ | $y$ |
|---|---|---|---|---|
| QSM | 1.70 ± 0.13 | 3.04 ± 0.01 | 1.05 ± 0.01 | 1.01 ± 0.01 |
| QSM_linear | **2.00 ± 0.08** | **3.90 ± 0.03** | **1.08 ± 0.03** | **1.03 ± 0.03** |

結果より，二次元空間，$x, y$ 軸から評価した場合においても QSM よりも QSM_linear のほうが行動分布の多峰性が高いことがわかる．

次に各モデルの行動分布を定性的に評価するために，Fig. 4 に同じエピソードにおいて開始から終了状態までの等間隔の状態において行動を 1,000 回生成した際のヒートマップを示す．

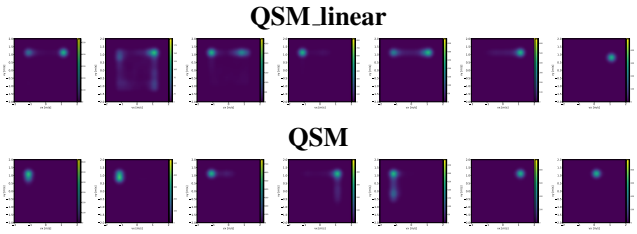


**Fig. 4** 2D action distribution heatmaps sampled at regular intervals within the same episode.

結果より，QSM_linear のほうが全体を通して行動分布が多峰性を持っていることがわかる．特に，エピソード序盤（左側）においては QSM では 1 つの行動に収束してしまっているのに対して，QSM_linear では行動が複数の箇所に集中していることがわかる．よって，アニーリングを行うことでモデルの多峰性を保持できることを確認した．また，どちらの手法もエピソード

終盤（右側）になると行動が一つに集中しているが，これはゴールに近づいていることによって行動がゴールへ向かう方向のみに集中しているためであると解釈できる．

### 4.4 条件付生成における性能評価

本研究では，$(x, y) = (-2, -5)$ から $(x, y) = (-2, 5)$，および $(x, y) = (2, -5)$ から $(x, y) = (2, 5)$ の位置に壁を 2 つ設置した．この環境では，前方から 3 人の歩行者が直進してくる状況を想定している．静的障害物を回避するように行動生成を式 (8),(9),(11) を用いて条件付けし，QSM と QSM_linear がガイダンススケール $\lambda$ に応じてどのように変化するかを比較した．その結果を Fig. 5 に示す．

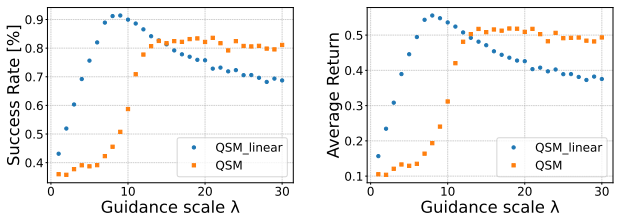


**Fig. 5** Success rate and average return as functions of guidance scale $\lambda$.

結果より，QSM よりも QSM_linear のほうがガイダンススケールを変化に対して成功率，平均収益のどちらの値も最大値に達するまでが早いことがわかる．また，お互いの最大値を比較した場合でも QSM_linear のほうが QSM よりも高い成功率，平均収益を獲得できていることがわかる．Table. 3 に各モデルの最も平均収益の高かったガイダンススケール値 $\lambda$ における SR（成功率），CWR（壁衝突率），CHR（歩行者衝突率），AR（平均収益）の比較を示す．

**Table 3** Comparison of QSM_linear and QSM at each method's optimal guidance scale for wall avoidance (mean ± std over 5 seeds).

| Method | SR [%] ↑ | CWR [%] ↓ | CHR [%] ↓ | AR ↑ |
|---|---|---|---|---|
| QSM | 0.84 ± 0.11 | **0.04 ± 0.03** | 0.06 ± 0.04 | 0.52 ± 0.09 |
| QSM_linear | **0.91 ± 0.10** | **0.04 ± 0.08** | **0.03 ± 0.01** | **0.55 ± 0.11** |

QSM よりも QSM_linear のほうが、すべての指標において同等か高いスコアを示した。これらの結果から、アニーリングを用いて拡散モデルを学習させることで、従来のモデルと比較して、ガイダンスを用いた条件付けによる行動生成時に壁回避と歩行者回避の両立が可能となり、性能を向上できることが確認された．

## 5. 実機実験

本説では，提案手法が実世界においても適用可能であるかを確認する．

### 5.1 実験環境

Fig. 6 に本実験で使用した 2 つの環境を示す．

1 つ目は Fig. 6(a) のような広い環境で歩行者回避実験に使用される．この環境ではロボットが通行することが想定される箇所には静的障害物が存在せず，歩行者のみを回避しゴールへ到達することができるかどうかを確認する．この実験においては歩行者を学習環境と同じように 5 人，円状に配置し対角線上に向かって移動するように指示している．

2 つ目は Fig. 6(b) のような横幅が 2 m の狭い通路環境で静的障害物回避実験に使用される．この環境ではロボットは歩行者と両サイドにある壁への衝突を回避しゴールへ到達する事ができるかどうかを確認する．この実験においては歩行者を 1 人ロボットの前に配置しロボットの方向に移動するように指示している．いずれの環境においてもモデルはシミレーション実験において多峰性，性能が最も高かった QSM_linear を使用し，ゴールはロボットから 8 m をとしている．


**(a)** Environment A **(b)** Environment B

**Fig. 6** Demonstration environments: (A) pedestrian avoidance, (B) static obstacle avoidance.

### 5.2 歩行者回避実験

歩行者回避実験では，ロボットが歩行者を認識し回避しながらゴールへ到達できるかを評価した．Fig. 7 にその実験結果を示す．

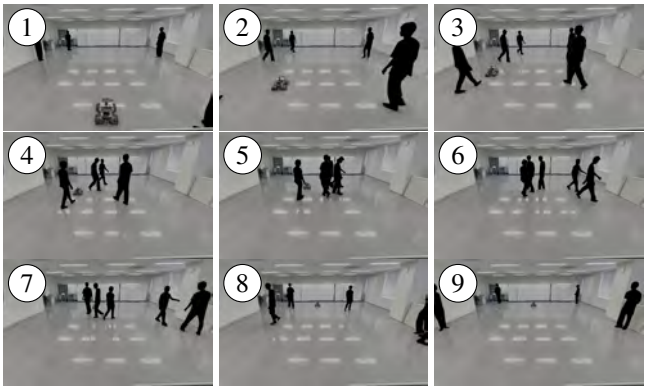


**Fig. 7** Scenes of the real-world demonstration using the proposed method in the circle-crossing scenario.

結果より歩行者を回避し，ゴールへ到達することを確認した．

### 5.3 静的障害物回避実験

静的障害物回避実験では，歩行者に加えて静的障害物が存在する環境での挙動を確認した．Fig. 8 にその実験結果を示す．

結果より，静的障害物が存在する環境においても歩行者と静的障害物の両方を回避しゴールへ到達することを確認した．


**Fig. 8** Scenes of the real-world demonstration using the proposed method in the corridor scenario.

## 6. まとめと今後の展望

本研究では，拡散モデルを強化学習の方策として導入し，逆温度パラメータ $\beta$ を学習進行に応じて制御する手法を提案した．これにより，方策モデルの性能が向上し，学習初期には探索を促進して学習効率を改善し，学習後期には方策の確率分布を収束させることでモード崩壊を抑制し，多峰性を保持できることを確認した．その結果，ガイダンスに基づく条件付き行動生成における性能が向上した．さらに，実機実験を通じて，提案手法が実世界においても適用可能であることを確認した．今後は，実環境での定量的な評価や適応性向上を目指して研究を進める予定である．

## 参考文献

[1] C. Chen, Y. Liu, S. Kreiss, and A. Alahi: Crowd-Robot Interaction: Crowd-aware Robot Navigation with Attention-based Deep Reinforcement Learning, Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2019), pp. 6015–6022.

[2] C. Chen, S. Hu, P. Nikdel, G. Mori, and M. Savva: Relational Graph Learning for Crowd Navigation, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2020), pp. 10007–10013.

[3] X. Zhang, W. Xi, X. Guo, Y. Fang, B. Wang, W. Liu, and J. Hao: Relational Navigation Learning in Continuous Action Space among Crowds, Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2021), pp. 3175–3181.

[4] Z. Wang, J. J. Hunt, and M. Zhou: Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning, Proceedings of the International Conference on Learning Representations (ICLR) (2023).

[5] J. Ho, A. Jain, and P. Abbeel: Denoising Diffusion Probabilistic Models, Advances in Neural Information Processing Systems (NeurIPS) (2020), pp. 6840–6851.

[6] 富田湧, 松本耕平, 兵頭侑樹, 倉爪 亮: 拡散モデルを利用した強化学習による歩行者混雑環境下での移動ロボットナビゲーション: 第 1 報, ロボティクス・メカトロニクス講演会 (2025), 2A1–O04.

[7] 富田湧, 松本耕平, 兵頭侑樹, 倉爪亮: 拡散モデルを利用した深層強化学習による歩行者混雑環境下での移動ロボットナビゲーション: 第 2 報, 日本ロボット学会学術講演会 (2025), 3E1–04.

[8] M. Psenka, A. Escontrela, P. Abbeel, and Y. Ma: Learning a Diffusion Model Policy from Rewards via Q-Score Matching, Proceedings of the International Conference on Machine Learnin (ICML) (2024), pp. 41163–41182.

[9] D. Rempe, Z. Luo, X. B. Peng, Y. Yuan, K. Kitani, K. Kreis, S. Fidler, and O. Litany: Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion, Proceedings of theIEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023), pp. 13756–13766.

[10] Y. Chen, C. Liu, B. E. Shi, and M. Liu: Robot Navigation in Crowds by Graph Convolutional Networks With Attention Learned From Human Gaze, IEEE Robotics and Automation Letters (RA-L), pp. 2754–2761 (2020).

[11] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha: Reciprocal n-Body Collision Avoidance, Proceedings of the International Symposium of Robotic Research (2011), pp. 3–19.

[12] B. W. Silverman: Using kernel density estimates to investigate multimodality, Journal of the Royal Statistical Society: Series B (Methodological), pp. 97–99 (1981).

[13] C. Fraley and A. E. Raftery: Model-based clustering, discriminant analysis and density estimation, Journal of the American Statistical Association, pp. 611–631 (2002).