토크-열 피드백 보상 설계를 통한 강화학습 기반 사족보행 안정성 향상

강재하

* 건국대학교 기계로봇자동차공학과

Improvement of Reinforcement Learning-Based Quadruped Locomotion Stability Using Torque-Thermal Feedback Reward Design

Jaeha Kang

* Dept. of Mechanical, Robotics and Automotive Engineering, Konkuk Univ.

Key Words: Quadruped Locomotion(사족보행), Reinforcement Learning(강화학습), Thermal Feedback(열 피드백), Torque Regularization(토크 정규화), Reward Design(보상 설계)

초록: 본 논문에서는 강화학습 기반 사족보행에서 온도 피드백 보상 설계가 보행 안정성에 미치는 영향을 분석하고, 토크 부하를 함께 고려한 토크-열 피드백 보상 구조를 제안한다. Baseline, Thermal Feedback, Thermal-Torque Feedback 정책을 Unitree Go2/MuJoCo 환경의 1.5 m/s 조건에서 비교하였다. 단순 온도 피드백 정책은 평균 속도는 증가하였으나 10 m 직진 시 lateral drift와 yaw drift가 크게 증가하여 방향 안정성이 저하되었다. 반면 토크-열 피드백 정책은 final lateral drift 0.065 m, final yaw drift 0.49 deg로 가장 안정적인 직진 보행을 보였다. 이를 통해 사족보행의 열적 안정성 개선에는 온도 상태뿐 아니라 발열 원인인 토크/부하를 함께 반영한 보상 설계가 필요함을 확인하였다.

Abstract: This paper analyzes the effect of thermal feedback reward design on reinforcement learning-based quadruped locomotion stability and proposes a torque-thermal feedback reward structure that also considers actuator torque load. Three policies, Baseline, Thermal Feedback, and Thermal-Torque Feedback, were compared in a Unitree Go2/MuJoCo environment under a 1.5 m/s command. The Thermal Feedback policy increased the average speed, but it also produced large lateral and yaw drift in the 10 m straight-walking task, resulting in degraded directional stability. In contrast, the Thermal-Torque Feedback policy achieved the most stable straight walking, with a final lateral drift of 0.065 m and a final yaw drift of 0.49 deg. These results show that thermal-aware quadruped locomotion requires reward design that reflects not only the temperature state but also the torque/load that causes motor heating.

Table 1 Nomenclature used in this paper

Symbol	Description
tau_i	i번째 관절에 작용하는 토크
omega_i	i번째 관절의 각속도
T_i	i번째 모터 온도
P_heat,i	i번째 모터의 열 입력
P_load	로봇 구동에 필요한 전체 부하 전력
SOC	배터리 충전 상태
R_th	열저항
tau_th	열 시정수
r_total	강화학습 전체 보상
r_locomotion	보행 성능 보상
r_thermal	열 상태 보상
r_torque	토크 또는 부하 집중 패널티

1. 서 론

사족보행 로봇은 다양한 지형에서 높은 이동성을 확보할 수 있어 점검, 운송, 탐사 등의 분야에서 활용 가능성이 크다. 최근에는 강화학습을 이용하여 복잡한 보행 정책을 학습하는 연구가 활발히 진행되고 있으며, 명령 속도 추종, 자세 안정성, 보행 리듬 등을 보상함수로 설계하여 안정적인 locomotion policy를 얻는 방식이 널리 사용된다. 그러나 고속 또는 장시간 보행 상황에서는 특정 관절 구동기에 토크 부하가 집중될 수 있으며, 이는 모터 손실과 온도 상승으로 이어져 보행 성능과 안정성에 영향을 줄 수 있다.

열 문제를 고려하기 위한 직관적인 방법은 motor temperature를 관측값 또는 보상함수에 포함하는 것이다. 그러나 온도는 토크 부하와 모터 손실이 누적된 결과값이므로, 단순히 온도 상태만을 반영한 정책은 발열 원인을 직접 제어하지 못할 수 있다. 특히 제한된 observation history를 사용하는 강화학습 정책에서는 장기적으로 누적되는 thermal state나 좌우 부하 불균형이 충분히 드러나지 않을 수 있으며, 이로 인해 특정 actuator에 torque demand가 집중되는 비대칭 보행이 학습될 가능성이 있다.

본 논문에서는 이러한 문제를 확인하기 위해 Baseline, Thermal Feedback, Thermal-Torque Feedback의 세 정책을 비교하였다. Baseline은 기본 보행 성능을 기준으로 하는 정책이며, Thermal Feedback은 온도 상태를 반영한 정책이다. Thermal-Torque Feedback은 온도뿐만 아니라 발열의 주요 원인인 토크/부하를 함께 고려하도록 설계하였다. Unitree Go2/MuJoCo 환경에서 1.5 m/s 조건의 보행 실험을 수행하고, lateral drift, yaw drift, energy per meter, motor temperature rise를 비교하여 토크-열 피드백 보상 설계가 보행 안정성에 미치는 영향을 분석하였다.

2. 강화학습 기반 보행 정책 및 보상 설계

2.1 강화학습 기반 사족보행 정책

강화학습 기반 사족보행 정책은 로봇의 현재 상태를 관측값으로 입력받아 각 관절에 대한 action을 출력하고, 이에 따른 보상의 누적값을 최대화하도록 학습된다. 본 연구에서 사용한 observation은 명령 속도, body/IMU 상태, 관절 위치와 속도, 이전 action 등으로 구성된다. 보상함수는 명령 속도 추종, 자세 안정성, 자연스러운 보행 리듬, 낮은 effort, smooth action 등을 포함하여 기본적인 locomotion 성능을 유도한다.

그러나 이러한 기본 보상만으로는 장시간 또는 고속 보행 중 특정 관절에 발생하는 지속적인 토크 부하와 온도 상승을 직접적으로 제어하기 어렵다. 따라서 본 연구에서는 기본 보행 정책을 기준선으로 두고, 온도 상태를 반영한 정책과 토크 부하까지 함께 반영한 정책을 비교하였다.

2.2 비교 정책 구성

본 연구에서는 Baseline, Thermal Feedback, Thermal-Torque Feedback의 세 정책을 비교하였다. Baseline 정책은 속도 추종과 자세 안정성을 중심으로 학습되며 motor temperature를 직접 관측하지 않는다. Thermal Feedback 정책은 motor temperature, thermal margin, hotspot 정보를 추가로 사용하여 온도 상승을 억제하도록 설계하였다. Thermal-Torque Feedback 정책은 온도 상태뿐 아니라 발열의 원인이 되는 torque/load 부담을 함께 고려하여 고온 actuator에 큰 토크가 집중되는 현상을 억제하도록 설계하였다.

여기서 중요한 차이는 Thermal Feedback이 온도라는 결과값에 반응하는 정책인 반면, Thermal-Torque Feedback은 온도를 발생시키는 원인까지 보상함수에 포함한다는 점이다. 이 차이를 통해 단순한 온도 피드백과 물리적 원인을 고려한 보상 설계가 보행 안정성에 미치는 영향을 비교할 수 있다.

Table 2 Comparison of reinforcement learning policies

Policy	Main feedback	Role	Expected limitation
Baseline	Locomotion state	기본 보행 기준선	열 상태를 직접 고려하지 않음
Thermal Feedback	Temperature state	온도 상승 억제	온도는 결과값이므로 부하 원인을 직접 제어하기 어려움
Thermal-Torque Feedback	Temperature + torque/load	발열 원인과 결과를 함께 제어	보상 가중치 설계가 필요함

2.3 보상 설계

세 정책은 공통적으로 속도 추종, 자세 유지, 발 움직임, 종료 패널티와 같은 기본 locomotion reward를 사용한다. Thermal Feedback 정책은 여기에 actuator 온도 변화량을 억제하는 thermal reward를 추가한다. Thermal-Torque Feedback 정책은 추가로 고온 actuator에 큰 torque가 집중되는 것을 억제하는 항과, 온도가 높을수록 torque 사용 여유를 줄이는 항을 포함한다.

전체 보상은 다음과 같은 형태로 정리할 수 있다.

$$r_{total} = r_{locomotion} + \lambda_T r_{thermal} - \lambda_{\tau} r_{torque}$$

여기서 $r_{locomotion}$은 기본 보행 성능 보상, $r_{thermal}$은 온도 상승 억제 보상, r_{torque}는 과도한 토크 또는 부하 집중에 대한 패널티이다. Baseline은 $r_{locomotion}$만 사용하고, Thermal Feedback은 $r_{thermal}$을 추가하며, Thermal-Torque Feedback은 $r_{thermal}$과 r_{torque}를 함께 사용한다. 이를 통해 thermal reward가 단순히 온도 상태에 반응하는 것이 아니라, 안정적인 locomotion objective와 정렬되도록 설계하였다.

3. 데이터 수집 및 열-부하 해석 모델

3.1 실로봇 데이터 수집 및 전처리

본 연구에서는 강화학습 정책의 열적 특성을 해석하기 위해 Unitree Go2의 sim-to-real 보행 과정에서 수집한 low-level 데이터를 사용하였다. 로봇의 상태와 명령은 ROS bag 형태로 저장하였으며, 주요 topic은 /lowstate와 /lowcmd이다. 원본 raw rosbag directory는 총 30개였고, 이 중 분석에 사용할 수 있도록 policy joint order에 맞추어 변환한 low-state CSV는 24개이다.

각 데이터는 약 10분에서 50분 동안 수집되었으며, 분석에는 실제 보행이 이루어진 active walking 구간을 중심으로 10분 window를 사용하였다. 최종적으로 20개의 train window와 4개의 test window를 구성하였고, 속도 조건은 0.5-1.5 m/s 범위의 실험 metadata를 기준으로 정리하였다. 이러한 전처리는 실제 모터 순서와 강화학습 policy에서 사용하는 joint order를 일치시키고, 토크, 각속도, 온도, 배터리 상태를 동일한 시간축에서 비교하기 위해 수행하였다.

Table 3 Telemetry summary used for thermal-load analysis

Item	Value	Description
Raw rosbag directories	30	sim-to-real walking logs
Low-state CSVs	24	/lowstate, /lowcmd converted to policy joint order
Analysis windows	20 train + 4 test	active walking windows
Velocity labels	0.5-1.5 m/s	experiment metadata labels
Motor telemetry	48 channels	12 motors x q, dq, tau, T
Battery/state telemetry	4 channels	voltage, current, SOC-related values

3.2 사용 데이터 항목

사용한 데이터 항목은 총 52개로 구성된다. 12개 모터에 대해 관절 위치 q, 관절 속도 dq, 추정 토크 tau_est, motor temperature를 사용하여 48개 motor telemetry를 구성하였다. 여기에 battery voltage, battery current, BMS SOC 관련 항목을 추가하여 전체 구동 전력과 배터리 상태를 함께 해석할 수 있도록 하였다.

특히 본 연구에서는 모터 온도를 단순한 관측값으로만 사용하지 않고, 토크 및 각속도에 의해 발생하는 손실과 연결하여 해석하였다. 이는 Thermal Feedback 정책이 온도라는 결과값에 반응하는 것과 달리, Thermal-Torque Feedback 정책이 발열 원인인 torque/load를 직접 보상 설계에 반영한다는 점을 설명하기 위한 기반이 된다.

3.3 열-부하 해석 모델

시뮬레이션에서 직접 얻을 수 있는 값은 각 관절의 토크와 각속도이다. 그러나 보행 중 발생하는 열 부담과 배터리 에너지 사용량을 해석하기 위해서는 토크, 각속도, 모터 발열, 모터 온도, 배터리 부하 사이의 관계를 모델링할 필요가 있다. 본 연구에서는 기존 물리 모델을 기반으로 한 grey-box 열-부하 모델을 사용하였다.

모터의 열 입력은 토크 제곱 기반의 구리손실, 속도 기반 마찰손실, Coulomb-type 손실, 그리고 모터 그룹별 bias 발열을 합산하여 다음과 같이 나타낼 수 있다.

$$P_{heat,i} = k_{tau,25}\{1 + \alpha_{Cu}(T_i - T_{ref})\}\tau_i^2 + b_v\omega_i^2 + \tau_c|\omega_i| + P_{bias,g(i)}$$

여기서 τ_i와 ω_i는 각각 i번째 관절의 토크와 각속도이며, T_i는 해당 모터 온도이다. $g(i)$는 i번째 모터가 속한 motor group을 의미하며, 본 연구에서는 Hip, Thigh, Calf 그룹으로 나누어 열 특성을 fitting하였다.

계산된 열 입력은 1차 RC thermal model을 통해 다음 시점의 모터 온도로 변환된다.

$$T_i(t+dt) = T_{amb} + \exp(-dt/\tau_{th,g(i)})[T_i(t) - T_{amb}] + R_{th,g(i)}\{1 - \exp(-dt/\tau_{th,g(i)})\}P_{heat,i}$$

또한 각 관절의 양의 기계 출력과 손실을 합산하여 배터리가 공급해야 하는 전체 부하 전력을 계산하였다.

$$P_{load} = P_{base} + \sum_i[\max(\tau_i\omega_i, 0)/\eta_{drive} + k_{tau,25}\tau_i^2 + b_v\omega_i^2 + \tau_c|\omega_i|]$$

이 모델의 목적은 모터 내부 권선 온도를 정밀하게 예측하는 것보다, 정책 평가에서 torque/load가 motor temperature rise와 energy consumption으로 연결되는 경로를 해석하는 데 있다. 따라서 본 연구에서는 해당 모델을 통해 단순한 온도 피드백보다 토크/부하를 함께 고려한 보상 설계가 필요한 이유를 물리적으로 설명하였다.

4. 실험 조건 및 평가 지표

4.1 실험 조건

본 연구에서는 세 가지 강화학습 정책의 보행 안정성과 열적 부담을 비교하기 위해 Unitree Go2 모델을 사용한 MuJoCo 기반 시뮬레이션 평가를 수행하였다. 비교 대상은 Baseline, Thermal Feedback, Thermal-Torque Feedback 정책이며, 모든 정책은 동일한 명령 속도 조건에서 평가하였다. 주요 명령 속도는 vx = 1.5 m/s로 설정하였으며, 고속 보행 상황에서 torque/load 분포와 방향 안정성이 어떻게 달라지는지 확인하였다.

정책 평가는 크게 두 가지 실험으로 구성하였다. 첫째, 20분 보행 실험을 통해 장시간 보행 중 이동거리, 평균 속도, 최대 모터 온도, 이동 거리당 에너지 사용량을 비교하였다. 둘째, 10 m 직진 명령 실험을 통해 lateral drift와 yaw drift를 측정하여 직진 안정성을 평가하였다. 또한 후속 분석에서는 480 s MuJoCo 평가를 통해 국소 열위험을 나타내는 thermal-risk 보조 지표를 별도로 확인하였다.

Table 4 Experimental conditions

Item	Setting
Robot / Simulator	Unitree Go2 / MuJoCo
Command speed	vx = 1.5 m/s
Compared policies	Baseline, Thermal Feedback, Thermal-Torque Feedback
Long walking task	20 min walking
Straight walking task	10 m straight command
Supplementary thermal-risk task	480 s MuJoCo evaluation

4.2 보행 성능 및 에너지 지표

20분 보행 실험에서는 각 정책이 동일한 속도 명령에서 얼마나 안정적으로 장시간 보행을 유지하는지 확인하였다. 평가 지표로는 이동거리, 평균 속도, 최대 모터 온도, energy per meter를 사용하였다. 이동거리와 평균 속도는 명령 속도 추종 성능을 나타내며, 최대 모터 온도는 보행 중 가장 높은 열 부담을 받은 actuator의 상태를 나타낸다. Energy per meter는 단위 이동거리당 필요한 구동 부담을 의미하므로, 보행 효율을 비교하는 지표로 사용하였다.

이 지표들은 단순히 빠르게 이동하는 정책이 좋은 정책인지 판단하기 위해 사용한 것이 아니다. 고속 보행에서 속도, 에너지 사용량, 모터 온도는 서로 trade-off 관계를 가질 수 있으므로, 본 연구에서는 평균 속도와 함께 열적 부담 및 에너지 효율을 함께 해석하였다.

4.3 직진 안정성 지표

10 m 직진 명령 실험에서는 정책이 목표 방향을 얼마나 잘 유지하는지 평가하였다. 이를 위해 10 m 도달 시점의 lateral drift와 yaw drift, 종료 시점의 final lateral abs와 final yaw abs를 사용하였다. Lateral drift는 로봇이 목표 직선 경로에서 좌우로 벗어난 정도를 나타내며, yaw drift는 heading 방향이 얼마나 틀어졌는지를 나타낸다.

특히 본 연구의 핵심 관심사는 thermal feedback이 보행 안정성을 해치지 않으면서 열 부담을 줄일 수 있는지 여부이다. 따라서 Thermal Feedback 정책이 높은 속도를 달성하더라도 lateral drift나 yaw drift가 증가한다면, 이는 안정적인 보행 개선으로 보기 어렵다. 반대로 Thermal-Torque Feedback 정책이 온도와 torque/load를 함께 고려하여 drift를 줄인다면, 이는 보상 설계가 locomotion stability와 더 잘 정렬되었음을 의미한다.

Table 5 Evaluation metrics for straight walking stability

Metric	Description	Purpose
10 m lateral drift	10 m 도달 시 좌우 이탈량	직선 경로 유지 성능
10 m yaw drift	10 m 도달 시 heading 오차	방향 안정성
Final lateral abs	종료 시 lateral drift 절댓값	최종 위치 오차
Final yaw abs	종료 시 yaw drift 절댓값	최종 방향 오차
Energy per meter	이동 거리당 구동 부담	보행 효율
Tmax rise	최대 모터 온도 상승량	열 부담

4.4 국소 열위험 보조 지표

발표 이후 추가 분석에서는 480 s MuJoCo 평가를 통해 국소 열위험 보조 지표를 확인하였다. 해당 지표는 20분 보행 실험 및 10 m 직진성 실험과 평가 길이와 metric 정의가 다르므로, 본 연구에서는 주 결과와 직접 평균내지 않고 보조 근거로 분리하여 해석하였다.

보조 지표로는 corrected thermal dose per meter, peak reported-temperature rise per meter, hotspot dose per meter를 사용하였다. Corrected thermal dose per meter는 이동거리당 누적 열 부담을 나타내며, peak reported-temperature rise per meter는 최대 보고 온도 상승을 이동거리 기준으로 정규화한 값이다. Hotspot dose per meter는 특정 actuator에 열 부담이 집중되는 정도를 나타낸다. 이 지표들은 Thermal-Torque Feedback이 단순 평균 온도뿐 아니라 국소적인 열 집중을 줄이는 데 기여하는지 확인하기 위한 목적으로 사용하였다.

5. 결과 및 해석

5.1 20분 보행 결과

Baseline 정책은 1542 m를 이동하였고, 평균 속도는 1.285 m/s, 최대 모터 온도는 71 deg C, energy per meter는 48 W/m로 나타났다. Thermal Feedback 정책은 1656 m로 가장 긴 이동거리와 1.379 m/s의 평균 속도를 보였으나, 최대 모터 온도는 95 deg C, energy per meter는 63 W/m로 증가하였다. 반면 Thermal-Torque Feedback 정책은 1612 m를 이동하여 Baseline보다 긴 이동거리를 보였고, 평균 속도도 1.343 m/s로 향상되었다. 동시에 최대 모터 온도는 67 deg C, energy per meter는 45 W/m로 세 정책 중 가장 낮게 나타났다.

정량 결과는 Table 6에 정리하였다.

Table 6 Twenty-minute walking metrics at command speed 1.5 m/s

Policy	Metric	Value
Baseline	Distance [m]	1542
Baseline	Mean speed [m/s]	1.285
Baseline	Max motor temp. [deg C]	71
Baseline	Energy/m [W/m]	48
Thermal Feedback	Distance [m]	1656
Thermal Feedback	Mean speed [m/s]	1.379
Thermal Feedback	Max motor temp. [deg C]	95
Thermal Feedback	Energy/m [W/m]	63
Thermal-Torque Feedback	Distance [m]	1612
Thermal-Torque Feedback	Mean speed [m/s]	1.343
Thermal-Torque Feedback	Max motor temp. [deg C]	67
Thermal-Torque Feedback	Energy/m [W/m]	45

위 결과에서 중요한 점은 높은 평균 속도가 반드시 좋은 보행 성능을 의미하지 않는다는 것이다. Thermal Feedback 정책은 가장 빠르게 이동했지만, 열 부담과 에너지 사용량이 크게 증가하였다. 이는 온도 상태만을 보상에 포함하는 방식이 actuator load를 균형 있게 분산시키지 못할 수 있음을 보여준다.

5.2 10 m 직진성 결과

Baseline 정책은 lateral drift +0.502 m, yaw drift +2.73 deg를 보였고, final lateral abs와 final yaw abs는 각각 0.515 m, 2.92 deg였다. Thermal Feedback 정책은 lateral drift -1.255 m, yaw drift -18.22 deg로 방향 안정성이 크게 저하되었으며, final lateral abs와 final yaw abs도 각각 1.311 m, 18.21 deg로 증가하였다. 반면 Thermal-Torque Feedback 정책은 lateral drift -0.065 m, yaw drift -0.34 deg를 보였고, final lateral abs와 final yaw abs는 각각 0.065 m, 0.49 deg로 가장 작았다.

정량 결과는 Table 7에 정리하였다.

Table 7 Straight-walking stability check under a 10 m command

Policy	Metric	Value
Baseline	Lateral drift [m]	+0.502
Baseline	Yaw drift [deg]	+2.73
Baseline	Final lateral abs [m]	0.515
Baseline	Final yaw abs [deg]	2.92
Baseline	Energy/m [W/m]	46.50
Baseline	Tmax rise [deg C]	3.50
Thermal Feedback	Lateral drift [m]	-1.255
Thermal Feedback	Yaw drift [deg]	-18.22
Thermal Feedback	Final lateral abs [m]	1.311
Thermal Feedback	Final yaw abs [deg]	18.21
Thermal Feedback	Energy/m [W/m]	68.56
Thermal Feedback	Tmax rise [deg C]	11.09
Thermal-Torque Feedback	Lateral drift [m]	-0.065
Thermal-Torque Feedback	Yaw drift [deg]	-0.34
Thermal-Torque Feedback	Final lateral abs [m]	0.065
Thermal-Torque Feedback	Final yaw abs [deg]	0.49
Thermal-Torque Feedback	Energy/m [W/m]	49.62
Thermal-Torque Feedback	Tmax rise [deg C]	3.90

위 결과는 본 연구의 핵심 결론을 잘 보여준다. Thermal Feedback은 온도 상태를 반영했음에도 불구하고 보행 방향이 크게 틀어졌으며, 에너지 사용량과 온도 상승도 증가하였다. 반면 Thermal-Torque Feedback은 온도와 함께 torque/load 부담을 고려함으로써 가장 작은 drift를 보였다. 따라서 열적 안정성을 고려한 보행 정책에서는 온도라는 결과값뿐 아니라 발열의 원인인 torque/load를 함께 제어해야 함을 확인할 수 있다.

시각화 결과에서 Fig. 1은 10 m 직진 명령에서 각 정책의 전체 진행 방향과 yaw 편차를 비교한 결과이며, Fig. 2는 동일한 직진 보행 조건에서 rear-foot trajectory와 body heading의 차이를 보여준다. Thermal Feedback에서는 발 궤적과 몸체 방향의 비대칭성이 크게 나타나는 반면, Thermal-Torque Feedback에서는 더 직선적인 발 궤적과 작은 heading 변화를 확인할 수 있다.

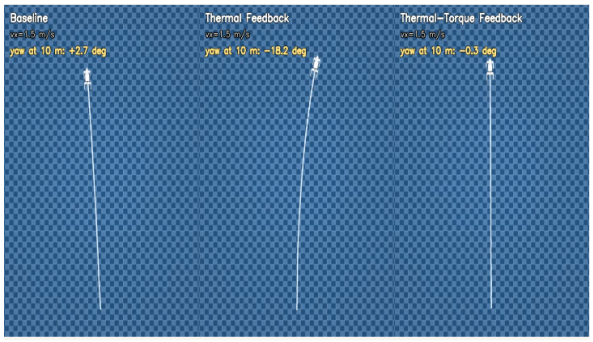

Fig. 1 Straight-walking yaw comparison at 10 m for Baseline, Thermal Feedback, and Thermal-Torque Feedback policies

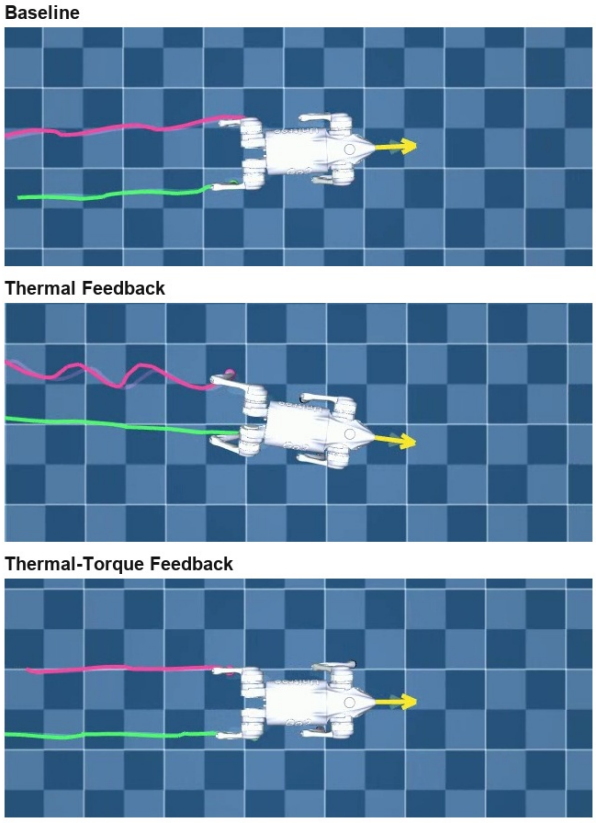

Fig. 2 Rear-foot trajectory and body heading comparison during straight-walking command at vx = 1.5 m/s

5.3 국소 열위험 보조 지표

후속 480 s MuJoCo thermal-risk 평가에서는 Thermal-Torque Feedback이 Baseline 대비 corrected thermal dose per meter, peak reported-temperature rise per meter, hotspot dose per meter를 각각 22.5%, 13.2%, 27.0% 감소시켰다. 이 결과는 10 m 직진성 결과와는 별도의 보조 evidence로, torque/load를 보상에 포함한 정책이 국소적인 열 집중을 줄이는 데에도 기여할 수 있음을 보여준다.

다만 이 지표는 20분 보행 실험 및 10 m 직진성 실험과 평가 길이 및 metric 정의가 다르다. 따라서 본 연구에서는 해당 결과를 주 결과와 직접 평균내지 않고, Thermal-Torque Feedback의 물리적 해석을 보강하는 보조 결과로 사용하였다.

5.4 MDP 관점의 해석

강화학습 기반 locomotion policy는 MDP 위에서 누적 reward를 최대화하도록 학습된다. 그러나 본 연구에서 사용한 policy는 LSTM과 같이 장기 내부 memory를 유지하는 구조가 아니며, control step마다 현재 observation과 제한된 history만으로 action을 결정한다. 이 경우 장기적으로 누적되는 actuator temperature, 특정 motor의 지속적인 부하 집중, 좌우 열 분포 불균형은 짧은 observation window에 충분히 드러나지 않을 수 있다.

Temperature-aware reward만 추가하면 정책은 전체 actuator의 열 부하를 균형 있게 낮추는 전략이 아니라, reward function의 빈틈을 이용해 특정 leg 또는 actuator에 torque demand를 집중시키는 전략을 학습할 수 있다. 이 경우 asymmetric torque allocation이 발생하고, gait symmetry가 무너지며, yaw drift와 lateral drift가 증가한다. 본 연구에서 Thermal Feedback 정책이 빠르지만 방향 안정성이 악화된 것은 이러한 failure mode와 일치한다.

Torque-aware regularization은 이러한 문제를 줄이는 physical regularizer로 작용한다. 과도한 torque 또는 current demand는 motor heating의 직접 원인이므로, torque/load margin을 penalty로 제한하면 temperature reward만으로 억제되지 않던 비정상적인 부하 집중을 줄일 수 있다. 결과적으로 Thermal-Torque Feedback은 온도 결과와 발열 원인을 함께 제어하여 yaw 방향 틀어짐을 줄이고, 직진 안정성을 유지하면서 열 부담을 낮추는 방향으로 학습되었다.

6. 결 론

본 논문에서는 강화학습 기반 사족보행에서 온도 피드백 보상 설계가 보행 안정성에 미치는 영향을 분석하고, 발열 원인인 토크/부하를 함께 고려한 토크-열 피드백 보상 구조를 제안하였다. 이를 위해 Baseline, Thermal Feedback, Thermal-Torque Feedback의 세 정책을 Unitree Go2/MuJoCo 환경에서 비교하였다.

실험 결과, 단순히 motor temperature를 반영한 Thermal Feedback 정책은 평균 속도는 증가하였으나 lateral drift, yaw drift, energy per meter, Tmax rise가 모두 증가하여 안정적인 보행 개선으로 이어지지 않았다. 이는 온도가 발열의 결과값이며, 제한된 observation history만으로는 장기적인 thermal state와 비대칭 torque allocation을 충분히 제어하기 어렵기 때문으로 해석된다.

반면 Thermal-Torque Feedback 정책은 온도 상태와 함께 torque/load 부담을 보상함수에 반영함으로써 10 m 직진 실험에서 final lateral drift 0.065 m, final yaw drift 0.49 deg로 가장 작은 방향 오차를 보였다. 또한 20분 보행 실험에서도 Thermal Feedback 대비 에너지 사용량과 최대 모터 온도를 낮추어, 보행 안정성과 열 부담 사이의 균형을 개선하였다. 따라서 사족보행 로봇의 thermal-aware locomotion을 위해서는 온도 상태를 단순히 추가하는 것보다, 발열을 유발하는 물리적 원인을 보상 설계에 함께 반영하는 것이 중요하다.

향후 연구에서는 동일한 초기 온도, 배터리 상태, 지면 조건, command sequence를 통제한 실로봇 장시간 반복 실험이 필요하다. 또한 장기 thermal state를 더 잘 반영하기 위해 recurrent policy 또는 thermal state estimator를 결합하는 방법을 검토할 수 있다.

후 기

본 연구는 2026학년도 종합설계 교과목의 일환으로 수행되었으며, 연구 진행 과정에서 지도와 조언을 주신 김남수 교수님께 감사드립니다.

참고문헌

(1) Wensing, P. M., Wang, A., Seok, S., Otten, D., Lang, J. and Kim, S., 2017, "Proprioceptive Actuator Design in the MIT Cheetah: Impact Mitigation and High-Bandwidth Physical Interaction for Dynamic Legged Robots," IEEE Transactions on Robotics, Vol. 33, No. 3, pp. 509-522.

(2) Wallscheid, O. and Bcker, J., 2016, "Global Identification of a Low-Order Lumped-Parameter Thermal Network for Permanent Magnet Synchronous Motors," IEEE Transactions on Energy Conversion, Vol. 31, No. 1, pp. 354-365.

(3) Lin, W., Qian, L., Luo, X. and Liang, C., 2025, "Temperature Distribution Prediction of the Quadruped Robot Based on the Lumped-Parameter Thermal Networks," Robot, Vol. 47, No. 2, pp. 188-199.

(4) Wang, Q., Gao, T. and Li, X., 2022, "SOC Estimation of Lithium-Ion Battery Based on Equivalent Circuit Model with Variable Parameters," Energies, Vol. 15, No. 16, 5829.

(5) Bernardi, D., Pawlikowski, E. and Newman, J., 1985, "A General Energy-Balance for Battery Systems," Journal of the Electrochemical Society, Vol. 132, No. 1, pp. 5-12.