

사족 보행 시각-언어-행동 모델에서의 코드북 붕괴 완회

김효빈, 김지원, 박태진, 이재구*

국민대학교 컴퓨터공학과

*jackoo@kookmin.ac.kr

요 약

코드북을 활용한 사족 보행 시각-언어-행동(Vision-Language-Action, VLA) 모델에서 코드북 붕괴(Codebook Collapse)로 인하여 로봇의 정밀한 제어를 방해하고 충돌과 같은 치명적 실패로 이어지는 문제가 발생한다. 기존 연구에서는 행동 청크 이산화(Action Chunk Discretization)에서 잔차 벡터 양자화(Residual Vector Quantization, RVQ)를 사용하였다. 이 과정에서 학습 중 다수의 코드가 사용되지 않는 코드북 붕괴가 발생하였다. 본 연구는 이를 극복하기 위해 양자화 전 레이어 정규화(Layer Normalization)와 코드 부활 기법을 RVQ 를 사용하는 사족 보행 시각-언어-행동 모델에 적용하였다. 실험 결과 제안 기법은 베이스라인 대비 코드북 활용도를 2.34%에서 99.20%로 크게 향상시켰으며, MAE 를 약 52.7% 감소시키고 PSNR 을 약 17.6% 높여 행동 복원 성능이 개선되었다.

1. 서론

최근 시각 정보와 언어 명령을 통합하여 로봇 제어를 수행하는 시각-언어-행동 (Vision-Language-Action, VLA) 모델이 활발히 연구되고 있다[1, 2, 3]. 기존 연구들은 로봇 팔을 이용하여 다양한 로봇 조작(Robotic Manipulation) 작업에서의 성능을 입증하였으며, 최근에는 이를 사족 보행 로봇으로 확장하려는 시도 또한 이어지고 있다[4, 5]. 일부 VLA 모델[3, 5]은 연속적인 로봇의 행동 데이터를 언어 모델이 처리 가능한 형태인 이산적 토큰으로 변환하기 위해, 잔차 벡터 양자화(Residual Vector Quantization, RVQ) 기법을 도입하여 사용한다.

그러나 RVQ 기반 VLA 모델은 기존 벡터 양자화(Vector Quantization, VQ) 방식과 마찬가지로, 특정 코드에만 편향되는 코드북 붕괴(Codebook Collapse) 문제에 직면해 있다[6]. 코드북 붕괴로 인해 코드의 상당 부분이 비활성화되면 잠재 공간의 표현 다양성이 제한되고, 복잡한 행동 패턴을 세밀하게 구분하는 변별력을 상실하게 된다. 이러한 제한된 표현력은 모델이 생성할 수 있는 유효 인덱스 조합을 축소시켜 재구성 품질 저하의 원인이 된다. 특히 미세한 복원 오차에도 민감하게 반응하는 로봇 제어의 특성상, 이러한 성능 저하는 치명적인 오동작으로 직결된다. 따라서 높은 복원 성능 확보는 정밀한 로봇 제어의 필수 요소이다.

이에 본 연구에서는 로봇 제어의 특수성을 고려하여, 코드북 붕괴 현상을 억제할 수 있는 코드북 부활 및 안정화 기법을 제안한다. 우리는 학습 초기 잠재 표현과 코드북의 정렬을 개선하는 초기화 전략, 학습 전반에 걸쳐 매칭 스케일의 안정성을 보장하는 정규화 기법, 그리고 소멸되는 코드를 되살리는 재활성화 기법을 통합적으로 설계하였다.

코드북을 사용하는 사족 보행 VLA 모델[5]에 제안 기법을 적용하여 행동에 대한 복원 성능 저하를 최소화하였다. 결과적으로 사족 보행 로봇의 안정적인 행동 생성 성능을 확보하였다.

2. 관련 연구

2.1 시각-언어-행동 모델과 행동 청크 이산화

QUART[4]는 주로 로봇 팔 조작 작업에 집중했던 기존 VLA 연구를 사족 보행 로봇으로 확장한 연구이다. QUART 는 지각(Perception)과 내비게이션(Navigation), 그리고 객체 회피와 같은 다양한 시나리오로 구성된 사족 보행 대규모 데이터 집합 QUARD[4]를 기반으로 학습함으로써 안정적인 동작 생성이 가능함을 보였다.

이후 제안된 QUART-Online[5]은 QUART 의 주요 한계였던 추론 지연 문제를 해결하기 위해 행동 청크 이산화 기법을 도입하였다. 연속적인 행동 시퀀스를 일정 길이의 청크로 묶어 매핑함으로써 단 순 이산화가 갖는 표현 중복성과 시간적 불연속성 문제를 완화하고, 긴 시퀀스를 효율적으로 처리할 수 있도록 하였다. 이를 통해 추론 지연 문제를 해결하여 효과적인 실시간 추론이 가능해졌다.

2.2 벡터 양자화 및 코드북 붕괴

연속적인 데이터를 이산적 코드로 매핑하는 벡터 양자화 기법은 VQ-VAE[7] 이후 생성 모델의 핵심 구성 요소로 자리 잡았다. 그러나 VQ 기반 모델은 학습 과정에서 인코더의 출력이 소수의 코드

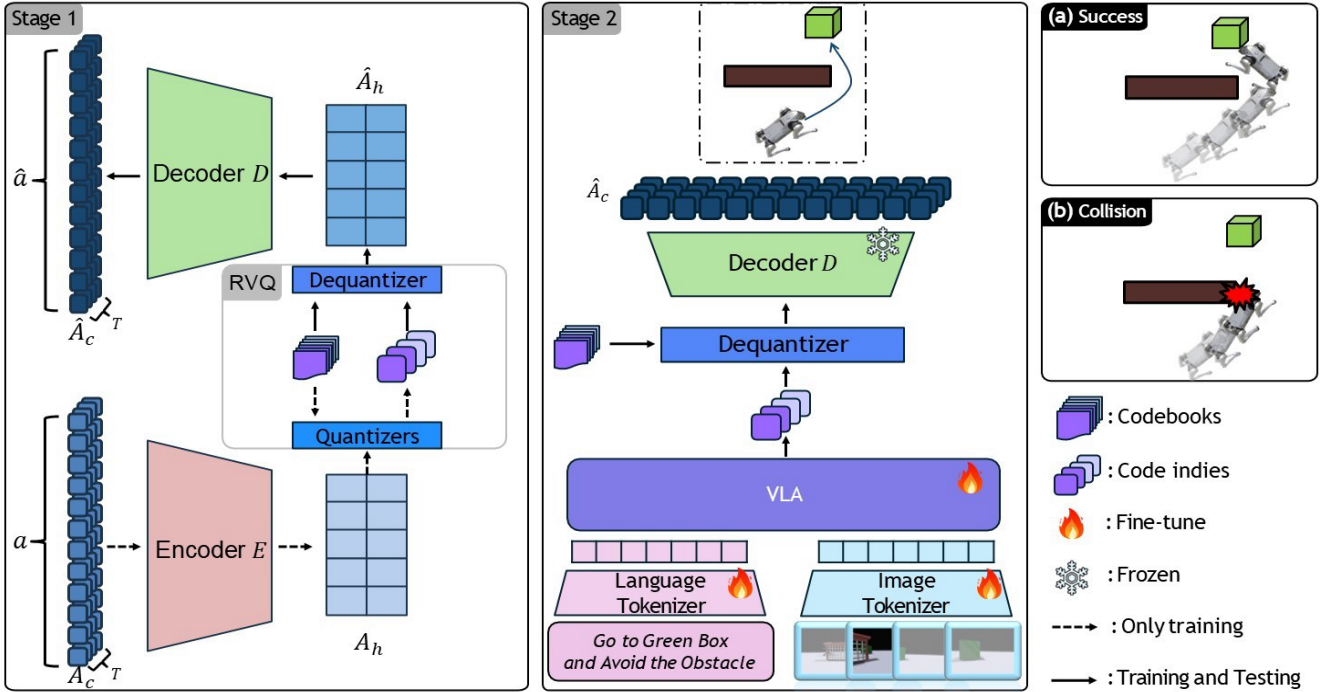


그림 1. QUART-Online 구조

에만 집중되는 코드북 붕괴 문제[8]에 취약하다. 코드북 붕괴가 발생하면 인코더가 입력 데이터의 다양성을 표현하지 못하고, 극히 일부 코드만을 반복적으로 사용하게 되어 재구성 품질을 크게 떨어뜨린다.

QUART-Online[5]에서 사용한 잔차 벡터 양자화 구조에서도 코드북 붕괴 문제가 발생한다. RVQ는 첫 번째 양자화기(Quantizer)가 데이터의 거시적인 특징을 잡고 이후 단계가 잔차를 줄여나가는 방식이다. 첫 번째 양자화기에서 붕괴가 발생하면 전반적인 특징을 놓친 채 과도한 오차를 후속 단계로 넘기게 되어 정보 전달 효율을 저하시킨다. 이는 데이터를 표현할 수 있는 유효한 인덱스 조합의 범위를 축소시켜 재구성 품질 저하로 이어진다. 이러한 복원 성능 저하는 정밀한 행동 제어를 방해하며, 결과적으로 로봇이 그림 1.(a)처럼 행동하지 않고 그림 1.(b)와 같이 충돌하는 상황이 발생할 수 있다.

2.3 코드북 부활 및 안정화 기법

이미지·오디오 생성 분야에서는 코드북 붕괴를 완화하기 위해 다양한 기법이 제안되어 왔으며 크게 두 가지로 분류할 수 있다. 첫 번째는 사용량이 임계값보다 낮은 코드를 임의로 재할당하는 무작위 재시작(Random Restart)을 사용함으로써 코드북 활용도를 높이는 방식[3, 9, 10]이다. 두 번째는 양자화 이전 단계에서 레이어 정규화(Layer Normalization, LN)를 적용하여 잠재 벡터들을 일정한 스케일로 유지함으로써, 유클리드 거리 기반 매칭의 안정성을 확보하는 방식[10]이다.

본 논문은 기존의 코드북 부활과 안정화 기법을 적용하여 로봇 제어에서의 코드북 붕괴 문제를 해결하고자 하였다.

3. 본론

본 연구의 목표는 사족 보행 로봇의 연속적인 행동 시퀀스를 손실 없이 이산화하여 정교한 행동을 생성하는 것이다. 이를 위해 QUART-Online[5]의 행동 청크 이산화 구조를 기반으로 하며, 학습 중 발생하는 코드북 붕괴 문제를 해결하기 위한 코드북 부활 및 안정화 기법을 제안한다.

3.1 배경: 베이스라인 모델 구조

QUART-Online에서 로봇의 행동 데이터는 시간 T 동안의 12 차원 연속적인 제어 명령 시퀀스로 구성된다. 입력 행동 청크 $A_c \in \mathbb{R}^{T \times 12}$ 는 전신 제어를 위한 12 차원의 물리적 제어 변수 집합으로 정의된다. 각 시점에서의 행동 벡터 a 는 수식 1과 같다.

$$a = [v_x, v_y, \omega_z, \theta_{1:3}, f, h_z, \phi, s_y, h_z^f, e] \quad (1)$$

여기서 v_x, v_y, ω_z 는 본체의 선속도 및 각속도, $\theta_{1:3}, f$ 는 보행 패턴 및 주파수, h_z, ϕ 는 본체 높이와 피치 각도, s_y, h_z^f 는 발의 너비와 높이, 그리고 e 는 행동 종료 신호를 나타낸다.

QUART-Online[5]의 구조는 그림 1과 같다. 먼저 그림 1의 Stage1에서 그림 1의 Stage2에 사용할 코드북과 디코더를 학습한다. 1D CNN 기반 인코더 E 로 행동 청크 A_c 를 잠재 표현 $A_h = E(A_c)$ 로 압축

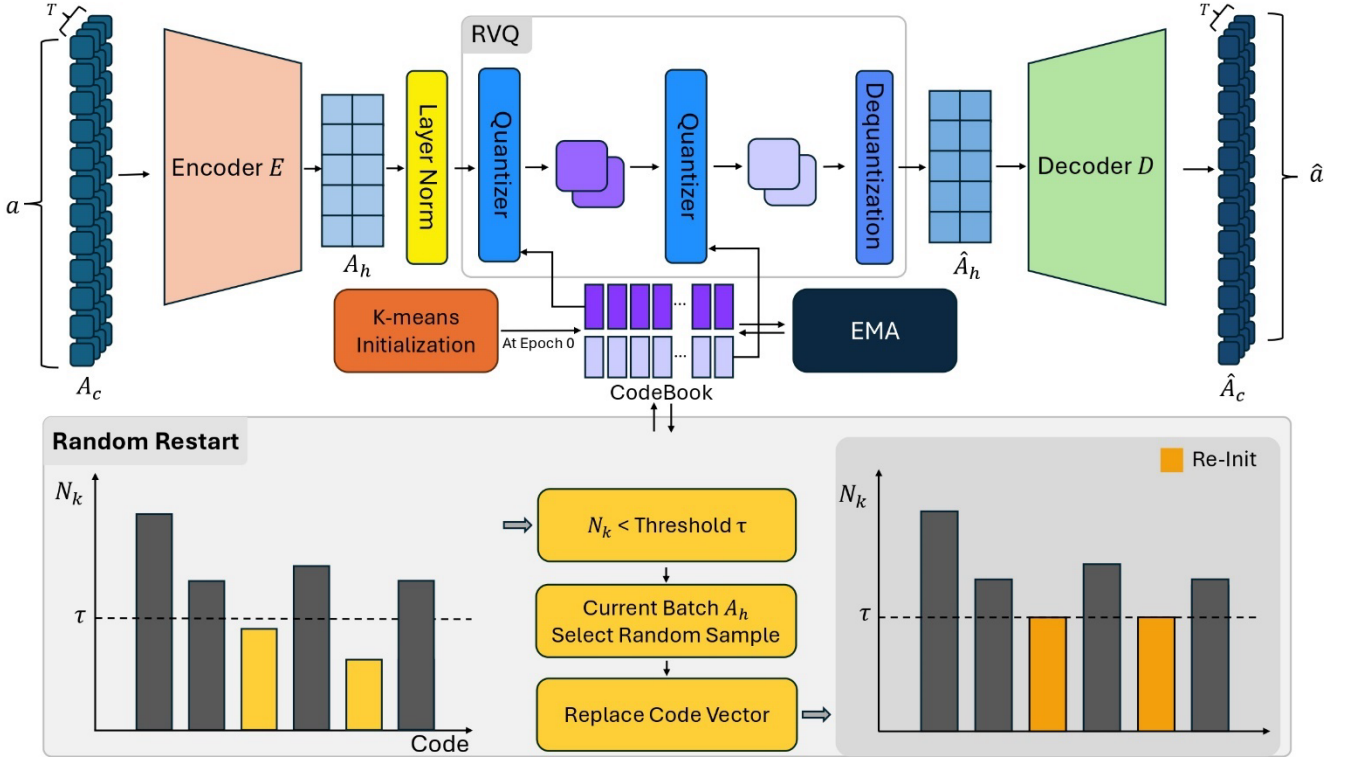


그림 2. 전체 구조도

한다. 이후 A_h 는 RVQ를 통해 이산화되고 각 단계 m 에서 잔차 r_{m-1} 는 코드북 C_m 내의 벡터 z_q^m 으로 근사된다. 즉, 코드북이 K 개의 벡터 집합 $C_m = \{c_k\}_{k=1}^K$ 로 구성될 때, 입력과 가장 유사한 특정 벡터 c_k 가 z_q^m 으로 선택된다. 최종적으로 선택된 코드들의 합 $\hat{A}_h$ 은 디코더 D 를 통해 $\hat{A}_c$ 로 복원된다.

학습 과정에서 코드북 벡터 c_k 는 지수 이동 평균을 통해 갱신된다. 각 코드 k 에 할당된 입력의 개수 n_k^s 와 입력 벡터들의 합 u_k^s 를 이용하여, 누적 사용량 N_k^s 와 누적 벡터 합 M_k^s 을 수식 2, 수식 3과 같이 갱신한다.

$$N_k^s = \gamma N_k^{s-1} + (1 - \gamma)n_k^s \quad (2)$$

$$M_k^s = \gamma M_k^{s-1} + (1 - \gamma)u_k^s \quad (3)$$

최종적으로 s 시점의 코드북 벡터 c_k^s 는 수식 4와 같이 계산된다.

$$c_k^s = \frac{M_k^s}{N_k^s} \quad (4)$$

그림 1의 Stage1에서 코드북과 디코더 학습이 완료되면 그림 1의 Stage2와 같이 VLA를 학습한다. 자연어 명령어와 이미지를 입력으로 넣어 이미지의 시점에 맞는 행동 청크의 코드 인덱스를 출력하도록 학습이 진행된다. 학습 이후 코드 인덱스는 앞에서 학습한 코드북과 디코더를 통해 행동 $\hat{A}_c$ 로 변환된다.

3.2 목적 함수

그림 1의 Stage1은 원본 행동의 재구성 품질을 향상시키고, 인코더 잠재 표현과 코드북 벡터 간의 정렬을 유도하기 위해, 재구성 손실(Reconstruction Loss)과 연계 손실(Commitment Loss)의 합을 최소화하도록 학습된다. 전체 목적 함수는 수식 5와 같다.

$$L_{\text{total}} = \|A_c - \hat{A}_c\|_2^2 + \|A_h - \text{sg}[\hat{A}_h]\|_2^2 \quad (5)$$

여기서 A_c 와 $\hat{A}_c$ 는 각각 원본 행동 청크와 복원된 행동 청크이며, A_h 와 $\hat{A}_h$ 는 인코더가 생성한 잠재 벡터와 양자화된 출력 벡터를 의미한다. sg는 Stop-gradient 연산으로, 역전파 과정에서 양자화 경로로 기울기가 흐르지 않도록 한다.

3.3 코드북 부활 및 안정화

QUART-Online은 RVQ 구조를 통해 효율적으로 행동 청크를 이산화하지만, 코드북이 붕괴하는 문제가 발생한다. 본 연구는 그림 1의 Stage1 학습 과정에서의 코드북 붕괴 문제를 해결하기 위해 양자화 이전 단계에 레이어 정규화를 적용하고, 학습 중 죽은 코드를 되살리는 무작위 재시작을 도입하였다. 또한 코드북을 초기화할 때 K-means 초기화 [10]를 사용하여 초기 학습 안정성을 확보하였다. 제안 방법의 전체 구조도는 그림 2와 같다.

표 1. 행동 복원 성능 및 코드북 활용도 비교

Metric Model	MSE(↓)	MAE (↓)	AKI (↓)	PSNR (↑)	UQI (↑)	Q_0 Utils- Batch (↑)	Q_1 Utils- Batch (↑)
QUART- Online[4]	0.00055	0.0112	0.00073	32.62	0.99991	2.34%	97.72%
Ours	0.00015	0.0053	0.00020	38.35	0.99995	99.20%	93.12%

3.3.1 양자화 전 정규화

인코더 출력의 스케일 변동은 거리 기반 선택의 안정성을 저해하여, 코드북 붕괴의 원인이 된다. 기존 QUART-Online[5]은 인코더 출력 A_h 를 별도의 정규화 없이 양자화기에 입력하므로, 배치 또는 시퀀스 간 스케일 변동에 민감해질 수 있다.

이를 완화하기 위해 본 논문에서는 인코더 출력 A_h 에 대하여 레이어 정규화를 수행하여, 잠재 벡터가 유클리드 거리 공간 상에서 보다 균일한 스케일을 갖도록 한다. 이를 통해 코드북 벡터와의 거리 비교가 더 공정하게 이루어지고, 학습 전반에 걸쳐 특정 코드에 대한 비정상적 집중 현상을 방지할 수 있다. 결과적으로 코드북의 안정성과 장기적 생존성을 모두 강화한다.

3.3.2 코드북 초기화 및 코드 부활

코드북 벡터들의 초기 분포가 인코더의 잠재 표현과 얼마나 잘 정렬되어 있는지는 학습의 안정성에 큰 영향을 미친다.

본 연구는 K-means 초기화[10]를 사용하여 코드북을 초기화하였다. 이 방식은 학습 시작 단계에서 첫 번째 배치의 잠재 벡터들에 대해 K-means 알고리즘을 수행하여 군집 중심점을 초기 코드북 벡터로 설정한다. 이는 인코더 출력과 코드북 벡터 사이의 초기 거리를 최소화하고, 모든 코드가 학습 초반부터 균등하게 선택받을 수 있는 기회를 제공함으로써 초기 학습 안정성을 확보할 수 있다. 학습 과정에서 죽은 코드(Dead code)가 발생하면, RVQ의 표현력과 이산화 효율이 크게 저하된다. 이를 해결하기 위해, 본 연구는 코드 부활 기법을 적용하였다. 코드 부활 기법으로 무작위 재시작을 사용하며, Algorithm 1과 같이 적용된다.

4. 실험

4.1 실험 환경 및 평가 지표

Algorithm 1 Codebook Update with Random Restart

Input Latent batch A_h ; Codebook $C = \{c_k\}_{k=1}^K$; Usage $\{N_k\}_{k=1}^K$; Vector sum $\{M_k\}_{k=1}^K$; Threshold τ ; Decay rate γ

```

1:  $A_h \leftarrow \text{LayerNorm}(A_h)$ 
2: for each latent vector  $a_i \in A_h$  do
3:    $k^* \leftarrow \text{argmin}_k \|a_i - c_k\|_2^2$ 
4:    $n_{k^*} \leftarrow n_{k^*} + 1$  and  $u_{k^*} \leftarrow u_{k^*} + a_i$ 
5: end for
6:  $N_k \leftarrow \gamma N_k + (1 - \gamma)n_k$ 
7:  $M_k \leftarrow \gamma M_k + (1 - \gamma)u_k$ 
8:  $c_k \leftarrow M_k / N_k$ 
9:  $K_{\text{dead}} \leftarrow \{k \mid N_k < \tau\}$ 
10: for each  $k \in K_{\text{dead}}$  do
11:    $a_{\text{rand}} \sim \text{Uniform}(A_h)$  // Sample random latent
12:    $c_k \leftarrow a_{\text{rand}}$ 
13:    $N_k \leftarrow \tau$ 
14:    $M_k \leftarrow c_k \times N_k$  // Re-sync vector sum
15: end for

```

Output Updated C , $\{N_k\}_{k=1}^K$, and $\{M_k\}_{k=1}^K$

본 연구는 사족 보행 로봇의 다중 과업 데이터 집합 QUARD[4]를 사용하여 제안한 코드북 부활 및 안정화 기법의 효과를 검증하였다. 입력 행동 청크의 길이는 $T = 10$ 으로 설정하였다. 비교 실험은 다음 두 가지 모델을 대상으로 수행하였다.

1. QUART-Online : 코드북 부활과 안정화 기법을 적용하지 않은 기본 RVQ 구조.
2. Ours : K-means 초기화, 레이어 정규화, 무작위 재시작을 적용한 기법.

행동 복원 성능은 MSE, MAE, PSNR, AKI, UQI로 평가하였으며, 코드북 붕괴가 발생하는지 측정하기 위해 코드북 활용도 (Utils)를 사용하였다. MSE(평균 제곱 오차)와 MAE(평균 절대 오차)는 수치적 복원 정확도를 나타내며, PSNR(최대 신호 대 잡음비)은 재구성된 신호의 품질을 측정한다. 구조적 관점의 유사성을 평가하기 위해 UQI(구조적 유사성)를 사용하였고, AKI를 통해 편향 및 분포 형태의 왜곡 정도를 분석하였다. 추가로 코드북 붕괴 현상을 확인하기 위해 Utils를 측정하였다.

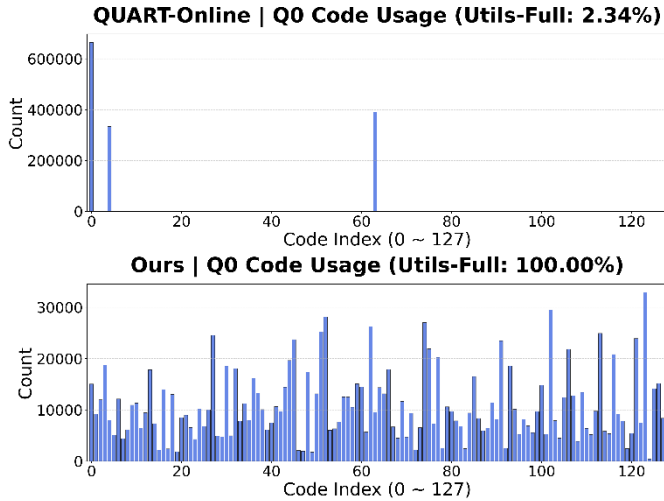
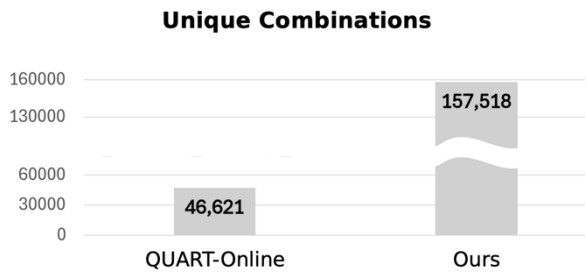
그림 3. Q_0 코드별 사용량 비교

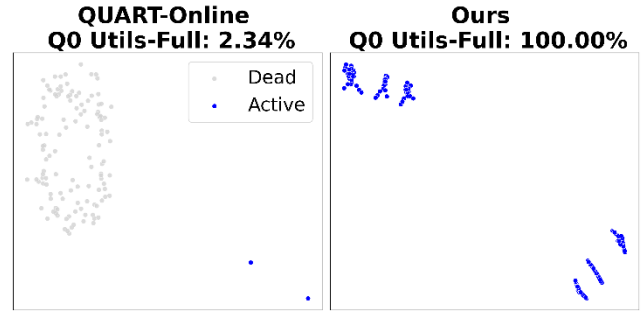
그림 4. 고유 인덱스 조합 수 비교

이때 첫 번째 양자화기 Q_0 Utils 와 두 번째 양자화기 Q_1 Utils 는 목적에 따라 두 가지 기준으로 측정된다. 첫 번째 기준(Utils-Batch)은 배치당 평균 활용도로 매 배치마다 사용된 코드의 비율을 평균 낸 값이다. 이는 모델이 단일 추론 단계에서 얼마나 풍부한 표현력을 가지는지 보여주는 지표이다. 두 번째 기준(Utils-Full)은 전역 누적 활용도로 검증 데이터 집합 전체에서 단 한 번이라도 선택된 적이 있는 코드의 비율을 의미한다. 이는 코드북 내에 영구적으로 죽은 코드가 존재하는지를 판별하는 지표이다. 지표별 성능에서 가장 높은 성능을 굵게 표시하였다.

4.2 행동 복원 성능 및 코드북 활용도 비교

표 1 과 그림 3 은 QUART-Online[5]이 심각한 코드북 붕괴 문제를 겪고 있음을 보여준다. QUART-Online의 첫 번째 양자화기 Q_0 의 Utils-Batch 와 Utils-Full 이 2.34%에 불과하다. 이는 128 개의 코드 중 약 3개만이 반복적으로 사용되었음을 의미하며, RVQ의 계층적 표현 학습이 작동하지 않았음을 시사한다. 이로 인해 MAE(0.0112), PSNR(32.62) 등 복원 성능이 크게 저하되었다.

반면, 제안된 Ours 는 코드북 붕괴를 근본적으로 억제한다. Ours 는 Q_0 Utils-Batch 를 99.20% , Utils-

그림 5. Q_0 코드북 t-SNE 분포 비교

Full 을 100%까지 올려 코드 사용 분포가 균일하게 확장되었음을 확인할 수 있다. 이는 무작위 재시작을 통해 죽은 코드를 즉시 부활시키고, 레이어 정규화를 통해 잠재 공간의 스케일을 안정화함으로써 재사망 문제를 방지한 결과이다. 그 결과 RVQ의 계층적 표현 학습이 원활히 이루어져 QUART-Online 대비 MAE는 약 52.7% 감소하였고 PSNR은 약 17.6% 증가하였다.

이러한 코드북 활용도의 개선은 모델이 생성할 수 있는 행동 표현의 다양성으로 직결된다. 그림 4에서 나타나듯, Ours 가 생성한 고유 인덱스 조합의 수는 157,518 개로, QUART-Online 의 46,621 개 대비 약 3.38 배 증가하였다. 이는 코드북 붕괴를 해결함으로써 더욱 정밀하고 다양한 사족 보행 행동 패턴을 생성할 수 있게 되었음을 보여준다.

4.3 코드북 분포 분석

그림 5 는 각 모델의 t-SNE 기반 코드북의 전역 누적 활용도를 시각화한 결과이다. 파란색 점은 검증 과정에서 한 번 이상 사용된 활성 코드이고, 회색 점은 완전히 선택되지 않은 죽은 코드이다.

QUART-Online[5]의 분포는 거의 전체가 죽은 코드로 채워져 있으며, 이는 전역 활용도가 97% 이상 붕괴했음을 의미한다. 이러한 현상은 QUART-Online 이 RVQ의 계층적 구조가 정상적으로 작동하지 못하고, 소수 코드에 잠재 표현이 집중되는 구조적 실패를 보여준다.

반면 제안 모델은 코드북 전체가 활성 코드로 구성되며, 특히 Q_0 계층에서 뚜렷한 군집 구조가 형성됨을 확인할 수 있다. 이는 RVQ 가 설계 의도대로 점진적 정밀화 방식(Coarse to fine)의 계층적 표현 학습을 수행하고 있음을 시각적으로 입증한다. 첫 번째 양자화기 Q_0 는 행동을 대략적으로 분리하는 다수의 안정적인 클러스터를 형성하였다. 이는 제안된 코드북 부활 및 안정화 기법이 단순히 붕괴를 막는 수준을 넘어 계층적 행동 표현 구조 자체를 활성화하는 핵심적 역할을 수행함을 보여준다.

5. 결론

본 연구에서는 사족 보행 로봇의 행동 생성을 위한 QUART-Online 모델에서 발생하는 코드북 붕괴 문제를 분석하고, 이를 해결하기 위한 코드북 부활 및 안정화 기법을 제안하였다. 실험 결과, 베이스라인 모델은 코드북 활용도가 2.34%에 불과하여 심각한 잠재 공간의 단편화 현상을 보였으나, 제안 방법은 코드북 활용도를 99.20%까지 향상시켜 학습 과정에서 안정적인 계층적 표현 학습을 가능하게 하였다.

특히 제안 방법은 불필요하게 세분화된 코드 인덱스 조합의 생성을 억제하여 학습에 필요한 데이터 밀도를 확보하였다. 이를 통해 베이스라인 대비 MAE 를 약 52.7% 감소시키고 PSNR 을 약 17.6% 향상시키며 우수한 행동 복원 성능을 입증하였다.

감사의 글

이 논문은 2025 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과임 (No.RS-2025-02219317, AI 스타펠로우십지원사업(국민대학교); No.RS-2025-02263754, 자율적 의사결정이 가능한 인간지향적 체화형 AI 에이전트)

참고문헌

- [1] Zitkovich, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." *Conference on Robot Learning*. PMLR, 2023.
- [2] Kim, et al. "Openvla: An open-source vision-language-action model." *arXiv preprint arXiv:2406.09246*, 2024.
- [3] Wang, et al. "VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers." *arXiv preprint arXiv:2507.01016*, 2025.
- [4] Ding, et al. "Quar-vla: Vision-language-action model for quadruped robots." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.
- [5] Tong, et al. "Quart-Online: Latency-Free Multimodal Large Language Model for Quadruped Robot Learning." *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.
- [6] Zheng, et al. "ERVQ: Enhanced residual vector quantization with intra-and-inter-codebook optimization for neural audio codecs." *IEEE Transactions on Audio, Speech and Language Processing*. 2025.
- [7] Van Den Oord, et al. "Neural discrete representation learning." *Advances in neural information processing systems* 30, 2017.
- [8] Zhu, et al. "Addressing representation collapse in vector quantized models with one linear layer." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025.
- [9] Dhariwal, et al. "Jukebox: A generative model for music." *arXiv preprint arXiv:2005.00341*, 2020.
- [10] Zeghidour, et al. "Soundstream: An end-to-end neural audio codec." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30, 2021: 495-507