SNU AI Challenge 방법론 보고서

비타민 Cv : 이지원, 장채린, 최연서, 한사랑

1. 전체적인 방법론 및 모델 구조

본 task 는 하나의 영상에서 추출된 4 장의 프레임이 무작위 순서로 제시될 때, 캡션과 시각적 단서를 활용해 원래의 시간 순서를 복원하는 문제이다. 4 장 모두 일치해야 정답으로 인정되는 Exact Match 기준이므로, 실질적으로는 24-way 순열 분류 문제에 해당한다.
우리는 Qwen3-VL-8B-Instruct 에 LoRA 를 적용해 fine-tuning 했다. 4 장을 동시에 비교해야 하는 특성을 고려해 언어 계층뿐 아니라 비전 인코더와 aligner 도 학습 대상에 포함시키되, 시각 계층에는 더 낮은 학습률을 부여해 사전학습된 표현의 붕괴를 막았다. 입력은 4 장을 Image 1~4 라벨과 interleave 해 배치하고 캡션과 지시문을 덧붙이는 형태이며, 지시문은 캡션에 순서 표현이 있으면 그것을 따르고 없으면 물체 위치나 이동 방향 같은 시각적 진행 단서로 판단하도록 두 경로를 분기시킨다.
설계 원리는 세 축으로 요약된다. 첫째, 입력 대칭성으로 학습 시 프레임 제시 순서를 4 가지로 증강하여 학습 데이터를 9059 건에서 36236 건으로 확장하였다. 둘째, 출력 공간 제약으로 자유 생성 대신 24rto 순열 후보의 log-likelihood 를 비교하는 listwise 추론을 적용하고, KV Cache 를 재사용하여 효율적으로 계산하였다. 셋째, 추론 분산 감소를 위해 서로 다른 입력 순서의 두 뷰(TTA)를 스코어링한 뒤 원본 프레임 기준으로 역매핑하여 합산하였다. 또한 학습, 추론, 제출 전 과정에서 단일 프레임 번호 변환 함수를 사용해 번호 불일치를 방지하였다.

2. 제안한 아이디어의 동기와 핵심 기여

이 과제는 출력 공간이 24 개(4!)뿐인 유한 분류 문제이고, 입력 프레임의 제시 순서가 임의로 부여된 좌표계일 뿐이라는 순열 대칭성이 내재하며, 난이도가 캡션의 순서 표지어 유무에 따라 이중 분포를 이룬다는 세 가지 성격을 갖는다. 본 접근의 핵심 기여는 이 중 '대칭성'을 학습과 추론 양쪽에서 명시적으로 활용한 것이다.

2.1 학습 : 순열 증강(K_AUG=4)

각 학습 샘플마다 입력 순열 σ 를 4 개(identity 1 + 랜덤 3) 생성해 프레임 배치와 정답(chrono 좌표)을 함께 변환, 별도 학습 행으로 기록한다(9,063→36,236 행). 증강이 없으면 모델이 "대체로 이미 정렬돼 있다"는 입력 순서 편향을 학습하게 되는데, "No_ordering=True" 케이스가 유의미한 비중을 차지하는 이 과제에서 이는 특히 위험하다.

identity 순열을 매번 고정 포함시킨 것은, 랜덤 순열만 쓰면 이 정당한 케이스의 분포가 희석되는 것을 막기 위한 의도적 설계다.

2.2 추론 : 순열 TTA (K_TTA=2)

입력 프레임 순서를 K 개 뷰로 바꿔 각각 24 개의 후보를 채점하고, 후보를 원본 좌표계로 역매핑해 로그우도를 합산한 뒤 argmax 를 취한다. 역매핑 없이는 서로 다른 좌표계의 점수를 더하는 셈이 되어서 TTS 가 오히려 해가 된다. 증강이 모델을 좌표계에 불변하게 만든다면, TTA 는 남은 좌표계 민감도를 평균으로 제거한다.
학습/추론이 하나의 설계 원리를 공유하는 지점이 되는 것으로 볼 수 있다.

3. 학습 및 추론 과정

3.1 데이터 구성

학습데이터 총 9539 개 샘플 중 validation set 은 seed 42 로 셔플한 후 상위 5%(476 개)를 고정하여 사용하였으며 나머지 9063 개를 학습에 사용하였다. 학습 데이터에는 4 배 순열 증강을 적용하여 최종적으로 36236 개의 학습 샘플을 생성하였다. Validation split 을 고정함으로써 백본과 학습 프레임워크가 변경되더라도 항상 동일한 476 개 샘플에서 성능을 비교할 수 있도록 하였다.

3.2 학습 설정

모델은 bf16 정밀도와 SDPA attention 을 사용하여 학습하였다. 총 1 epoch(2265 step)동안 학습하였으며, batch size 는 1 에 gradient accumulation 16 을 적용하였다. Learning rate 는 cosine scheduler 와 3% warmup 을 사용하였고 최대 입력 길이는 4096, 이미지 토큰 수는 128~320 으로 설정하였다. 또한 gradient checkpointing 을 적용하였으며, 200step 마다 체크포인트를 저장하고 최근 3 개만 유지하였다. 모든 실험은 seed 42 로 수행하였다.

3.3 추론 과정

추론은 자유 생성 대신 24 개(4!) 순열 후보에 대한 listwise 로그우도 스코어링으로 수행하였다. 각 순열을 후보 문자열로 구성한 뒤 후보 토큰의 로그우도 합을 계산하여 가장 높은 점수를 갖는 순열을 최종 예측으로 선택하였다. 이 방식은 invalid 출력을 방지하고 생성 기반 디코딩의 오류 누적을 완화한다. 계산 효율을 위해 이미지, 캡션 프리픽스를 한 번만 forward 하여 KV Cache 를 생성하고, 이를 모든 후보가 공유하도록 하여 뷰당 한 번의 이미지 인코딩으로 24 개 후보를 모두 평가하였다. 또한 추론 시 순열 기반 Test-Time Augmentation(TTA, K=2)를 적용하여 두 입력 순열 뷰의 로그우도를 원본 좌표계로 역매핑해 합산한 뒤 최종 순열을 선택하였다.

4. 성능 향상을 위해 적용한 주요 기법

4.1 문제 상황

기존 체크포인트의 오류를 캡션 길이별로 분해하자 (0,18] 단어 구간 EM 23.65%, (30,40] 구간 86.62%로 극단적으로 갈렸다. 모델은 프레임 간 시간 관계를 읽는 것이 아니라 캡션의 순서 표지어를 번역하고 있었다. 순열 증강이 이 의존을 학습 신호 차원에서 억제하는 방법론이라면, 이하의 기법들은 모델이 실제로 사용할 시각 경로의 해상력과 채점 정밀도를 끌어올리는 방법론이다.

4.2 VLM 의 시각 경로 강화

세 가지를 함께 조정했다. 첫째, 이미지당 비주얼 토큰을 최대 320 까지 올렸다. 순서 표지어가 없는 샘플은 미세한 위치 변화나 누적 변화량으로 판단해야 하는데, 저해상도에서는 4 장이 사실상 동일하게 인코딩되어 비교 자체가 성립하지 않는다. 다만 4 장 동시 입력이라 시퀀스 길이가 해상도에 비례해 늘어나므로 이전 실험에서 병목으로 확인된 지점까지만 올렸다. 둘째, 비전 인코더와 aligner 를 freeze 하지 않았다. 사전학습된 시각 표현은 "무엇이 있는가"를 인코딩할 뿐 "어느 쪽이 시간상 앞인가"를 구분하도록 학습된 적이 없기 때문이다. 다만 언어 계층과 같은 학습률로 갱신하면 표현이 붕괴하므로 1e-4 대 2e-5 로 차등을 두었다. 셋째, 학습 대상이 전체 모듈로 확장되면서 늘어난 적응 부담에 맞춰 LoRA rank 를 32 에서 64 로 상향했다.

이 조합은 전체 성능을 끌어올렸지만 정작 목표였던 (0,18] 구간은 25.0%로 거의 움직이지 않았다. 시각 표현의 해상력만으로 해결되지 않는 병목이 남아 있다는 뜻이며, 이는 6 장의 한계 논의로 이어진다.

4.3 후보 채점의 정밀도 향상과 비용 확보

listwise 스코어링은 후보 간 log-likelihood 를 공정하게 비교할 수 있을 때만 성립한다. 이를 위해 24 개 후보의 토큰 길이가 모두 동일한지(본 설정 13 토큰) 실행 시점에 검증한다. 길이가 다르면 짧은 후보가 합산에서 구조적으로 유리해지기 때문이다. 또한 종료 토큰까지 포함해 채점하는데, 이를 빼면 다른 순열의 접두사가 될 수 있는 후보가 부당하게 높은 점수를 받는다. 비용 측면에서는 프롬프트(이미지 4 장 포함, 1,000 토큰 이상)를 24 번 재계산하지 않도록, 프롬프트 구간을 한 번만 전파해 만든 KV 캐시를 24 배 복제하고 후보 토큰 13 개만 추가 전파했다. 실질 비용이 1 회 전파 수준으로 내려가면서 TTA 를 추가할 상황이 충족되었다고 할 수 있다

4.4 프롬프트의 조건부 분기

4.1 의 문제 상황에 대한 판단을 프롬프트에도 반영해, 캡션에 순서 표현이 있으면 그것을 따르고 없으면 물체 위치/이동 방향/누적 변화량 같은 시각적 진행 단서로 판단하도록 지시문 안에서 두 경로를 분기시켰다. 단일 지시문 내 조건부 기술이므로 어댑터를 나누거나 추론 시 라우팅하지 않으며 단일 모델 제약을 유지한다.

5. 실험 결과 및 분석

5.1 주요 결과

held-out val(476 개, seed 42) 기준 v5 최종 모델의 전체 및 세부 구간별 EM 은 다음과 같다.

전체 476 개 샘플에서 Exact Match(EM)는 0.6008 이었다. 이 중 순서 표지어가 있는 368 개 샘플의 EM 은 0.701, 순서 표지어가 없는 108 개 샘플의 EM 은 0.259 로 나타났다. 정답이 identity 인 경우의 EM 은 0.636, 비-identity 인 경우는 0.594 였다.

리더보드 test EM : 0.89. 증강 없는 baseline(r32, LLM 만 적응) 대비 val EM +3.2%p(0.569 → 0.6008)

단, rank·학습범위·TTA 가 동시에 바뀐 교란 비교다. identity/비-identity 격차(4%p)와 표지어 유무에 따른 성능 차이가 갖는 의미는 6 절(장단점)에서 다룬다.

5.2 분석

val(476 개)은 순서 표지어 보유율 77.3%, 45 단어 초과 캡션이 476 개 중 1 개인 반면, test 는 각각 89.5%, 29.4%로 test 가 val 보다 구조적으로 더 쉬운 분포를 갖는다. test 캡션 길이 분포는 0–15 단어 10.3%, 15–30 단어 28.4%, 30–45 단어 31.9%, 45 단어 초과 29.4%다.

val 에는 45 단어 초과 샘플이 사실상 없어, test 샘플의 29.4%가 속하는 구간을 val 로는 전혀 검증할 수 없다. val 의 구간별 EM 에 test 의 길이 분포를 곱해 재가중하면(0.103×0.291 + 0.284×0.579 + 0.319×0.868 = 0.471, 여기에 45+구간 기여분을 더함), 45+구간이 완벽(EM=1.0)하다고 가정해도 상한은 약 0.765 다. 실제 리더보드 test EM 0.89 는 이 상한을 12%p 이상 초과하는데, 이는 길이 재가중만으로는 val-test 격차가 다 설명되지 않으며 표지어 보유율 차이(89.5% vs 77.3%)처럼 길이와 독립적인 난이도 요인이 추가로 작용하고 있음을 시사한다.

캡션이 긴 샘플일수록 표지어를 포함해 텍스트만으로도 순서가 풀리는 경향이 있어, 긴 문장 서브셋의 EM 은 test 성능의 근사적인 참고 지표로 볼 수 있다. val EM(0.60)을 test

성능의 절대적 추정치로 쓰면 실제 성능을 크게 과소평가하며, val 에서 관측된 개선폭 역시 test 에서의 개선폭을 그대로 대표하지 못한다. 다만 남은 개선 여지 역시 결국 val 이 가장 취약한 그 짧은/표지어 없는 층에 있다고 볼 수 있다.

6. 방법론의 장점과 한계

6.1 장점

본 방법론의 강점은 유한 분류(listwise scoring)와 순열 대칭성 활용이라는 두가지 핵심 설계 원칙에서 비롯된다.

첫째, 순서 복원 문제를 24 개 순열 중 하나를 선택하는 24-way 분류 문제로 재정식화였다. 각 후보의 로그우도를 계산하여 최적의 순열을 선택하는 방식으로 invalid 출력과 파싱 실패를 구조적으로 제거하였으며, 모든 후보를 전역적으로 비교하여 생성 기반 디코딩의 누적 오류를 줄였다. 또한 의사결정 과정이 Exact Match 평가 지표와 직접 정렬 되며, 출력 형식을 고정하여 후보 간 로그우도를 고정하게 비교할 수 있도록 설계하였다.

둘째, 문제에 내재한 순열 대칭성을 학습과 추론에서 일관되게 활용하였다. 학습에서는 순열 증강을, 추론에서는 순열 기반 TTA 를 적용하여 입력 순서에 대한 불변성을 학습하고 남아 있는 순서 민감도를 완화하였다. 실제로 No_ordering=Ture 와 False 샘플의 Exact Match 차이가 4%p 이내로 나타나 모델이 입력 순서보다 내용 자체를 기반으로 판담함을 확인하였다.

또한 결정론적 추론과 고정된 TTA 를 통해 높은 재현성을 확보하였으며, 프롬프트와 라벨 규약을 통일하여 train-inference skew 를 최소화하였다. 아울러 KV Cache 재사용과 LoRA 기반 학습을 적용하여 단일 24GB GPU 에서도 효율적인 학습과 추론이 가능하도록 설계하였다.

6.2 한계

본 방법론의 가장 큰 한계는 시각적 시간 추론 능력이 제한적이라는 점이다. 순서 표지어가 없는 캡션에서 Exact Match 가 0.259 에 머물러 프레임 간의 시간적 변화를 직접 이해하기보다 캡션과 이미지의 대응 관계에 크게 의존함을 확인하였다. 또한 비전 인코더를 함께 학습한 경우에도 성능 향상이 제한적이었던 점은 병목이 시각 표현보다 문제 정식화나 데이터 특성에 있을 가능성을 시사한다. 또한 계산 자원의 제약으로 인해 요인별 ablation 을 충분히 수행하지 못하였다. 비교 실험에서 LoRA 설정, 순열 증강, 학습 범위, TTA 등이 동시에 변경되어 각 요소의 개별 기여도를 분리하여 분석하지 못했으며, 학습량과 체크포인트 선택 역시 제한된 범위에서만 탐색하였다.

방법론 자체의 확장성에도 한계가 있다. 현재의 로그우도 기반 스코어링은 각 순열을 독립적으로 평가하므로 순열 간의 구조적 유사성을 고려하지 못한다. 그 결과 정답과 유사한 후보들의 정보를 효과적으로 활용하지 못한다.