텍스트로 풀어보는 장면의 재구성 AI Challenge 결과보고서

순열 스코어링 기반 비디오 프레임 순서 복원

팀: 4Reel (정지은, 김민정, 김서영, 김선형)

Ⅰ. 서론 및 문제 정의

하나의 영상에서 추출되어 무작위로 섞인 4장의 프레임과 해당 영상을 설명하는 캡션이 주어질 때, 프레임의 시간 순서를 복원하는 과제이다. 출력은 입력 i번째 프레임이 정답 순서에서 차지하는 위치를 나타내는 역순열(inverse permutation) 형태이며, 평가 지표는 4개 순서가 모두 일치해야 정답으로 인정되는 Exact Match(EM)이다.

본 연구의 핵심 설계는 생성에서 분류로의 관점 전환이다. 4장의 순서는 4!=24가지로 유한하므로, 순서를 텍스트로 생성하는 문제 대신 24개 순열 후보 중 로그우도(log-likelihood)가 최대인 후보를 선택하는 분류 문제로 재정식화한다.

[파이프라인 개요]

백본: Qwen3-VL-8B-Instruct (멀티모달 VLM)
적응: LoRA(r=32, α=64) 파인튜닝, 비전 인코더 동결·언어모델 투영층만 학습
학습 신호: (캡션 + 4프레임) → 시간순 나열, 정답 토큰에 한정한 손실
편향 제어: 학습 시 프레임 순서 셔플 증강, 추론 시 자기일관성 TTA(self-consistency test-time augmentation)
디코딩: 24-순열 로그우도 스코어링 후 argmax
효율화: KV(prefix) 캐시를 통한 24후보 채점의 중복 연산 제거

[용어 정의]

본 보고서에서 leak-free val은 이미지 재사용 누수를 제거한 자체 검증셋을, label-free 지표는 정답 없이 모델 거동을 진단하는 지표군을 지칭한다.

[선행연구 근거]

Sort Story(EMNLP'16)는 캡션과 이미지가 상호보완적이며 텍스트 단서가 특히 강함을 보였고, OPN(ICCV'17)은 순서 예측을 순열 분류로 환원하는 정식화를, Made to Order(2024)는 트랜스포머 기반 정렬을 제시하였다. 본 연구는 이들을 참고하되, Sort Story가 순서를 생성 방식으로 접근하여 무효 출력 위험을 갖는 한계를 유한 후보 스코어링으로 해소한다.

Ⅱ. 제안 방법의 동기 및 기여

제안 방법은 세 가지 문제의식에서 출발하였다.

(i) 생성 방식은 중복([1,1,2,3])이나 누락([1,2,3]) 등 무효 출력을 낼 수 있어, 24-순열 스코어링은 항상 유효한 순열만을 출력하도록 보장한다.

(ii) 통상의 test-time augmentation(crop, 좌우 반전)은 시간 방향을 훼손하므로, 이미지 변형 대신 프레임 제시 순서를 셔플하는 태스크 특화 TTA를 설계하였다.

(iii) test 라벨이 없고 train에 이미지 재사용 누수가 있어, leak-free 검증셋과 label-free 지표 체계를 함께 구축하였다.

이는 Public 점수 과최적화보다 방법론의 견고성과 독창성을 우선한다는 판단에 따른 것으로, 이후 모든 실험은 점수 향상 자체가 아니라 관찰된 개선이 실제 신호인지 노이즈인지를 구별하는 데 초점을 둔다.

[핵심 기여]

1. 24-순열 로그우도 스코어링 디코딩 : 출력 유효성 보장 및 파싱 실패 제거
2. 프레임 순서 셔플 : 학습 증강과 추론 TTA를 통한 위치 편향의 이중 상쇄
3. train 이미지 재사용 구조의 규명 : 리더보드 점수와 실제 추론력의 괴리를 정량화하고 전량 학습 결정의 근거를 확보
4. KV 캐시 적용을 통한 예측 동일 조건에서의 추론 약 13.8배 가속
5. margin·시드 분산·Kendall·입력 ablation을 아우르는 label-free 진단 체계 : 근거 없는 변경으로 제출 기회를 소모하지 않는 의사결정 프로토콜의 정립

Ⅲ. 학습 및 추론

[학습 데이터]

대회 제공 train 9,535건만을 사용하며 외부 데이터·API는 사용하지 않았다. 5장에서 규명하듯 리더보드 성능이 train 이미지의 재사용 기억에 크게 의존하므로, 검증셋 확보를 위해 데이터를 남기는 것보다 전량 학습으로 커버리지

를 극대화하는 편이 리더보드 점수에 유리하다고 판단하여 홀드아웃 없는 전량 학습을 채택하였다. 다만 이는 오프라인 검증 가능성을 포기하는 트레이드오프를 수반한다.
아울러 초기에는 고빈도 재사용 이미지·전 프레임 저조도 세트를 학습에 부적합한 노이즈로 간주해 제거하는 방안을 검토하였으나, 재사용 상위 이미지의 94.7%가 실제로는 정상 밝기의 유효 콘텐츠였고 전 프레임 저조도 세트(41건, 0.43%) 역시 육안 검수 결과 대부분 정상적인 야간·저조도 장면으로 확인되어 정제를 보류하였다. test에도 동일한 분포의 프레임이 존재할 것으로 판단하여 데이터를 제거하지 않고 그대로 학습에 사용하였다.

[학습 설정]

LoRA(r=32, α=64, dropout=0.05)를 q/k/v/o 및 gate/up/down 투영층에 적용한다. 실효 배치 8(배치 1 × gradient accumulation 8), 학습률 1e-4(cosine, warmup), 해상도 768(비전 토큰 768개 상당의 픽셀 예산), 2에폭으로 학습하며, 100스텝마다 체크포인트를 저장하여 중단 시 자동 재개한다.

[프롬프트(학습·추론 공통)]

Storyline: "{캡션}"과 함께 4장(Image 1–4)을 제시하고, 스토리라인을 근거로 시간 순서를 시간순 이미지 번호 리스트로만 답하도록 지시한다.

[손실 마스킹]

정답 순열 토큰에만 cross-entropy를 부여하고 프롬프트·이미지 토큰은 -100으로 마스킹하여 순서 출력만을 학습한다.

[증강]

매 스텝 입력 4프레임 순서를 무작위로 셔플하고 정답 라벨을 동시 변환한다. 셔플 전후 "시간순으로 본 프레임 목록"의 보존을 단위 테스트로 검증한다.

[추론 절차]

1. 24개 순열 각각을 teacher-forcing으로 부착해 로그우도를 계산하고, 뷰별 정규화 없이 합산하여 argmax를 취한다.
2. 입력 프레임 순서를 무작위로 재배치한 뷰를 K=7회 추론하고, 각 결과를 원 좌표계로 되돌려 로그우도를 누적 합산한다.
3. 추론 해상도를 학습과 동일한 768로 정렬한다.
4. 예측 순열을 역순열로 변환(24순열 왕복 단위 테스트로 검증)하고 형식 검증(행 수·ID·순열 유효성·분포)을 거쳐 제출한다.

[재현성]

추론 루프 직전 random/numpy/torch 시드를 고정하여 TTA의 무작위 뷰 선택을 결정론화하며, 동일 코드·동일 가중치 재실행 시 동일 결과를 보장한다.

Ⅳ. 핵심 기법

(A) 24-순열 스코어링

4장의 순서 후보는 24가지로 유한하므로, 각 후보를 정답 문자열로 붙여 teacher-forcing 로그우도를 계산하고 최댓값을 택한다. 자유 생성과 달리 출력이 항상 유효한 순열이므로 중복이나 누락으로 인한 파싱 실패가 원천 차단되며, 1위와 2위의 점수 차(margin)를 확신도 지표로 곧바로 활용할 수 있다.

(B) 순서 셔플 증강 및 자기일관성 TTA

zero-shot 모델은 입력 슬롯 위치를 정답으로 답하는 위치 편향을 나타낸다. 학습 증강으로 이를 완화하고, 추론 시 다중 뷰를 원 좌표계로 되돌려 합산함으로써 잔여 편향을 상쇄하였다. 동일 모델에서 K=1 대비 K=7 적용 시 리더보드 점수가 0.883에서 0.904로 0.021 상승하여 TTA의 실효를 확인하였다.

(C) LoRA 파인튜닝

모델 크기 선택을 위해 train 1,000개 서브셋·1에폭의 예비 실험으로 2B와 7B를 동일 조건(r=32, 해상도, 증강)에서 비교하였다. train loss가 각각 0.211, 0.122로 7B가 뚜렷하게 낮았다. leak-free val EM도 0.17 대 0.24로 같은 방향이었으나 두 신뢰구간이 겹쳐(n=100) 통계적 확정은 아니며 최종적으로는 리더보드 점수(0.880 -> 0.932)가 8B 계열로의 스케일업의 근거가 되었다.

(D) KV(prefix) 캐시

24개 순열 후보를 채점할 때 공통으로 들어가는 이미지·프롬프트 구간의 Key/Value를 1회만 계산하여 재사용하였다. 언어모델의 인과 어텐션 구조상 뒤에 오는 답 토큰은 앞선 프리픽스의 표현에 영향을 주지 못하므로, 캐시를 재사용해도 각

후보를 독립적으로 채점한 것과 동일한 로그우도가 산출되었다. 최종 모델로 두 채점 경로를 같은 조건에서 대조한 결과 예측은 전부 일치하였고 추론 속도는 약 13.8배 빨라져, 예선 test 추론이 약 1.5시간에 완료된다. 이 절감을 통해 K=24 완전열거, 시드 11벌, 입력 ablation 등의 분석을 마감 내 수행할 수 있었다.

V. 실험 및 분석

최종 성적은 Public 0.93193, Private 0.91056 로 기록되었으며, 모델 스케일과 파인튜닝의 단조 기여가 관찰되었다. (2B zero-shot 0.106 → 2B QLoRA 0.72 → 8B 0.932)

모델	파인튜닝	설정	Private	Public
Qwen2-VL-2B	zero-shot	생성	0.09349	0.10645
Qwen2-VL-2B	QLoRA	24순열, K=7	0.75609	0.72425
Qwen2.5-VL-7B	LoRA	24순열, K=5	0.86585	0.87958
Qwen3-VL-8B	LoRA	24순열, K=1, 512	0.87398	0.88307
Qwen3-VL-8B	LoRA	24순열, K=7, 512	0.91869	0.90401
Qwen3-VL-8B	LoRA	24순열, K=7, 768	0.91056	0.93193
Qwen3-VL-8B	LoRA	K=24 완전열거	0.92276	0.91972
Qwen3-VL-8B	LoRA	균형 8뷰	0.92682	0.92321

표 1. 모델·설정별 성능 종합

(1) 자체 검증 점수가 리더보드와 벌어지는 원인

[관찰] train 9,535문제 × 4장 = 38,140슬롯 중 파일 내용(바이트 MD5) 기준 고유 이미지는 24,181개에 불과하다. 한 이미지가 최대 80개 문제에 등장하고, 48.8%의 문제가 타 문제와 이미지를 공유하며, 재사용 이미지의 정답 위치 최빈값 비율은 평균 0.649(무작위 기대치 0.25)에 달한다.

[해석] 모델이 [이미지→위치] 대응을 암기하는 것만으로 train 이미지를 재사용하는 test에서 고득점이 가능하다. leak-free val(0.23–0.29)과 리더보드(0.88–0.93)의 약 0.6 격차는 두 지표가 각각 순수 추론력과 이미지 재사용 이점을 측정하기 때문이다.

[함의] 분석 과정에서 test 정답을 직접 조사하거나 학습에 활용하지 않았으며, train 통계와 우리 모델의 예측 분포, 공개 점수만을 근거로 삼았다. 이는 서로 다른 세 모델(7B 태깅: val 0.2897 / LB 0.87958, 8B: val 약 0.23 / LB 0.9 이상, 7B 무태깅: val 0.2767)에서 동일한 방향으로 재현되어 독립적으로 뒷받침된다. 아울러 학습 데이터 암기 점검 셀에서 무작위 10건 중 10건이 모두 일치한 사례는 EM이 실제로 0.28 수준이라면 우연히 10건을 모두 맞힐 확률이 약 1/340,000에 불과해 통계적으로 사실상 불가능하며, 해당 표본이 학습 과정에서 암기된 결과임을 뒷받침한다.

(2) 지표 신뢰성 - 제로샷 캘리브레이션

제로샷 모델의 leak-free val EM은 0.114(95% CI 0.095–0.134)로, 동일 제로샷의 리더보드 점수(공식 베이스라인 0.136, 자체 0.106)와 동일 대역·동일 방향에 위치한다. 자체 점수 0.106이 신뢰구간 안에 들어와, 검증셋이 인위적으로 부풀려지거나 왜곡되지 않았음을 확인하였다. 자체 val 점수가 근소하게 낮게 측정된 것은 leak-free val 내 identity(순서 무의미) 정답 비율이 전체 대비 낮아 상대적으로 난이도가 높기 때문이며, 이는 검증셋을 인위적으로 부풀리지 않고 보수적으로 구성한 결과로 해석된다.

(3) TTA 뷰 수 분석

뷰를 몇 개까지 늘려야 하는지 확인하기 위해 K를 1부터 24까지 완전히 열거하였다. 24개 뷰를 모두 합산한 예측을 기준으로 삼으면 K=1은 93.3%, K=7은 95.8%가 일치한다. 일치율은 답이 얼마나 바뀌는지만 알려줄 뿐 어느 쪽이 더 정확한지는 말해주지 않으므로 리더보드로 확인하였다. K=1에서 K=7로 늘릴 때는 바뀐 예측이 대체로 옳은 방향이어서 점수가 0.021 상승하였다(4장-(B) 참조). 반면 K=7 이상에서는 개선이 없었다. K=7(0.932), 균형 8뷰(0.923), K=24(0.920) 순으로 오히려 소폭 낮아졌고, 특히 K=7과 K=24의 차이는 Public에서 K=7 우세(+0.0122), Private에서 K=24 우세(+0.0122)로 부호가 대칭 반전되었다. 이는 실제 성능 차가 아니라 노이즈로 판단된다.(시드 노이즈 ±0.0066, 표준오차 ±0.011과 같은 크기) 이에 최종 K를 7로 확정하였다.

(4) 시드 분산 - 노이즈 정량화

시드만 변경한 11회 반복에서 예측 변동률은 3.57%±0.42%, 만장일치 비율은 90.4%였다. 변동 샘플은 서로 상쇄

되므로 EM 변동폭은 ±0.0066으로 추정된다.(이론상 최댓값은 ±0.0357) 따라서 소규모 점수차는 방법의 우열이 아닌 시드 분산에 기인한다.

(5) 확신도(margin) 분석 - 노이즈 단일 원천 규명

24개 순열 중 1·2위 로그우도 차(margin, K=24 뷰 합산)를 확신도 프록시로 정의한다. 뷰를 합산하면 margin도 뷰 수에 비례해 커지므로 여기서의 수치는 모두 K=24 뷰 합산 기준이다. margin 6 초과 구간(780건)의 뷰 간 순서 불변율은 79.6%인 반면 6 이하(39건)는 0%이며, 시드 간 예측이 변동한 49건은 평균 margin 8.0으로 불변 샘플(123.5)의 1/15이고 전부 margin 하위 25%에 속한다. 즉 시드 분산과 뷰 불일치, 저확신이 대체로 같은 소수 집합에 몰려 있다. 뷰 간 답이 갈린 198건 중 87%가 margin 하위 25%에 속한다. 모델은 예측의 75.8%를 뷰와 시드에 무관하게 동일하게 산출하며, 실행 간 노이즈는 나머지 소수 샘플에 국한된다.

(6) Identity 편향 분석 및 캡션 서사 기반 교정 실험

test 후반부에서 모델이 입력 순서 그대로인 [1,2,3,4]를 연속 출력하는 구간이 관찰되었다. 확신이 없을 때 모델이 입력 순서로 회피한다는 가설을 세우고, 캡션의 순서 지시어가 3개 이상이면서 1, 2위 로그우도 차(단일 패스 기준)가 작은 샘플에 가중치를 주는 보정 로직을 설계하였다. 그러나 정답이 있는 train 300건으로 검증한 결과 가설은 반증되었다. 저margin 구간에서 identity로 답한 샘플의 정답률이 오히려 더 높았고(0.632 vs 0.518), identity 답변은 저margin보다 고margin 구간에서 더 자주 나타났다(20.4% vs 8.3%). 모델의 identity 예측 비율(13.3%)이 train의 identity 정답 비율(15.5%)와 근접한 것도 같은 방향의 근거이다. identity 예측은 모델이 판단을 피한 결과가 아니라 학습 데이터에서 가장 흔한 정답을 따른 것이었다. 교정할 대상이 아니었으므로 보정 로직은 채택하지 않았다. 규칙 기반 교정이 통하지 않는 더 근본적인 원인은 (7)에서 규명한다.

(7) 입력 ablation과 실패 샘플 분석

캡션 전체 제거 시 예측이 69.1% 변화하고 확신도가 90% 감소한 반면, 순서 단어(then/after/finally)만 제거한 경우 2.6%만 변화하였다. 이는 모델이 순서 연결어가 아니라 각 장면의 설명 내용과 이미지를 대응시켜 순서를 결정함을 시사하며, 규칙 기반 [STEP] 태깅이 무효였던 원인(강조 대상인 연결어를 모델이 근거로 사용하지 않음)을 설명한다. 한편 4장을 모두 동일한 단색 이미지로 대체한 경우 86.7%가 변화하였는데, 이 조건에서는 24개 순열의 로그우도가 수학적으로 같아져 항등 순열이 강제 선택된다. 따라서 이는 이미지 의존도의 측정이라기보다 시각 입력이 소실됐을 때의 퇴화 조건을 확인한 것에 가깝다. margin 하위 80샘플은 캡션이 평균 19.8단어로 상위 80샘플(33.4단어)보다 짧았고, 시간 전환어를 포함한 비율도 51%로 96%에 크게 못 미쳤다. 상당수가 캡션과 영상 모두에 순서 근거가 없는 본질적 모호 데이터였다.

(8) Kendall(쌍별 순서) 분석

EM은 4장 전부 일치를 요구하여 부분 정확도를 반영하지 못한다. 4프레임의 6개 쌍별 순서 일치를 측정하면 뷰 간 완전일치는 75.8%이나 쌍별 일치는 97.2%이다. 정답이 존재하는 7B val 50건에서도 EM은 0.14인 반면 평균 오답 쌍은 6쌍 중 2.4쌍에 불과하다. 즉 EM은 모델의 순서 이해도를 과소평가하며, 오답 대부분은 인접 한 쌍의 뒤집힘이다.

(9) 집계 방식 분석

softmax(T=0.3) 집계는 7B leak-free val paired test(n=400)에서 raw_sum 대비 ΔEM +0.0175(부트스트랩 CI 하한 >0)로 유의하였으나, McNemar 검정 p=0.065로 통상 유의수준에 미달하였고 8B 리더보드에서는 역전(0.914 < 0.923)되었다. 이에 최종 집계는 단순 합(raw_sum)을 유지한다.

별도로 온도 T=3.0의 softmax 집계도 시험하였으나, 확신도가 과도하게 평탄화되며 정답 순열의 상대적 우위가 희석되어 오답을 택하는 사례가 관찰되었다. 두 온도 조건(T=0.3, T=3.0) 모두 raw_sum을 넘지 못해, 최종 집계는 raw_sum으로 확정하였다.

(10) 효과 없던 시도 (Negative Results)

가설·결과·기각 근거를 기록하여 탐색의 충실성을 제시한다.

시도	가설	결과	판정
캡션 [STEP] 교정	저신뢰 시 identity 회피 교정	저margin(단일 패스 기준)에서 identity 정답률 0.632 vs 0.518	기각(합리적 사전확률)
규칙 기반 [STEP] 태깅	캡션의 순서 연결어를 명시 마킹해 순서 학습 강화	ΔEM +0.013 (CI [-0.012, +0.039], p=0.364)	기각(통계적 중립), 위 (7)에서 원인 규명
SimPO 선호학습	near-miss 억제로 EM 개선	val 0.290 → 0.267	기각

LOFO 오류정정	3장 부분순서 합의로 보정	ΔEM −0.034 (OOD)	기각(3장 입력이 학습 분포 밖)
consensus 집계	이상 뷰 감쇠	ΔEM −0.008 (p=0.65)	기각
균형 8뷰(라틴스퀘어)	저K 분산 감소	0.923 < 0.931	기각(Public 기준, Private에서는 역전, 6장 참조)
로그우도 정규화 합산	뷰 편향 보정	성능 하락	기각
DoRA 파인튜닝	크기·방향 분해로 LoRA 대비 표현력 향상	크기 벡터 연산 오버헤드로 학습시간·VRAM 증가, 성능 이득 미미	기각(효율 대비 이득 부족)
Pairwise 랭킹 추론	2장씩 비교하는 방식이 4장 동시 추론보다 유리	val 100건 기준 4장 동시 추론과 차이 없음	기각(중립)
프롬프트 엔지니어링	장면 묘사 주목 지시로 개선	ΔEM -0.009 (CI [-0.020, +0.002], p=0.163)	기각

프롬프트 변형은 학습을 거치지 않은 제로샷 모델로 지시문 한 문장을 추가해 비교하였다. 파인튜닝 모델은 학습 프롬프트에 맞춰져 있어 직접 비교 시 프롬프트 효과와 학습-추론 불일치가 뒤섞이기 때문이다. 1,001건 중 968건(96.7%)에서 예측이 동일하였고(ΔEM −0.009, CI 포함 0), 이는 모델이 캡션 장면 묘사에 의존한다는 (7)의 결론과 일관된다. 다만 제로샷 모델은 점수가 바닥권(val 0.114, 항등 베이스라인 0.155)이라 효과가 있더라도 작게 나타날 수 있다. 따라서 이 결과가 파인튜닝 모델에서의 무효까지 보장하지는 않는다.

VI. 논의: 장점과 한계

[장점]

24-순열 스코어링은 생성 방식의 무효 출력과 파싱 실패를 구조적으로 제거한다. 셔플 증강과 TTA는 위치 편향을 학습 및 추론 양쪽에서 이중으로 상쇄한다. KV 캐시는 예측을 바꾸지 않으면서 추론을 약 13.8배 단축하여 단일 RTX 3090에서 예선 test를 약 1.5시간에 처리하며, 24시간 규정에 충분한 여유가 있다. 나아가 정답 없이 모델을 진단하는 label-free 지표 체계와 노이즈 인지 평가를 통해 소규모 점수차에 현혹되지 않고 근거 기반으로 제출을 결정하였다.

[한계]

1. TTA가 4!=24의 유한 순열에 의존하여 프레임 수 증가 시 순열이 계승적으로 증가하는 확장성 제약이 있다.
2. leak-free val로는 리더보드를 예측할 수 없어(5장-(1)) 전량학습으로 전환하였고 그 결과 Public 점수 외에 신뢰 가능한 오프라인 지표가 없었다. 실제로 균형8뷰(Private 0.92682/Public 0.92321)는 최종 채택된 K=7(Public 0.93193/Private 0.91056)보다 Private에서 더 높았으나 Public만으로는 이를 구분할 수 없었다. 다만 이 차이 역시 5장-(4)에서 측정한 노이즈 범위와 겹치므로 문제의 본질은 특정 모델을 놓친 것보다 어떤 선택이든 근거를 댈 수단이 없었다는 데 있다. 향후에는 전량 학습본과 검증용 홀드아웃 모델을 함께 유지해 선택 근거를 확보할 필요가 있다. 그러나 5장-(9)에서 보았듯이 leak-free val의 상대 순위조차 리더보드에서는 뒤집혔으므로 홀드아웃만으로 이 문제가 풀린다고 보기는 어렵다. 오프라인 지표 하나로 리더보드를 대체하기보다 홀드아웃과 label-free 진단을 교차 검증하는 설계가 필요하며, 본 보고서에서 서술한 label-free 진단이 그 한 축이 될 수 있다.
3. 성능의 상당 부분이 이미지 재사용 기억에 기인하여 신규 영상에 대한 일반화(leak-free 0.2대)가 낮으며, 이는 태스크·데이터 구조에서 비롯한 본질적 상한이기도 하다.
4. 5-(6)의 margin 진단은 train 무작위 표본(n=300, 저margin×identity n=19)으로 수행되어 통계적 검정력이 제한적이며, 가설의 출처였던 test 후반부 고해상도·저조도 서브도메인을 충분히 대표하지 못했을 가능성이 있다. 아울러 일부 검증 실험은 4bit 양자화 상태에서 평가되어 최종 bf16 추론과 정밀도가 달랐으며, 그 차이가 비교 결과에 미친 영향은 별도로 확인하지 않았다. 뷰별 로그우도를 정규화 없이 합산하는 방식이 전제하는 뷰 간 독립성도 검증 대상에서 제외되었다.

VII. 자원 효율 및 구축 비용

최종 추론은 LoRA 어댑터를 병합한 모델을 bf16으로 실행한다.(양자화 미적용) 실측 VRAM 사용량은 약 18GB로 단일 RTX 3090(24GB) 내에서 동작한다. 예선 test(819샘플) 기준 추론 시간은 약 1.5시간(6.6초/샘플), 외부 검증 데이터(3,884 샘플) 기준 3.63시간(3.37초/샘플)이다. 외부 데이터는 이미지 해상도가 낮아 샘플당 처리 시간이 더 짧다. 학습은 외부 API·추가 데이터 없이 Google Colab 및 개인 GPU만으로 구축하였다.