텍스트로 풀어보는 장면의 재구성

SNU AI Challenge 2026 기술 보고서 · 팀 태셔 · 2026년 7월

Public LB	Private LB
0.93368 (15위)	**0.92682** (7위)

1 서론

본 과제는 한 문장의 캡션과 무작위로 섞인 정지 프레임 4장을 입력받아 원래의 시간 순서를 복원하는 문제이다. 후보 공간은 $4! = 24$개의 순열로 유한하며, 평가는 시퀀스 전체가 정답과 완전히 일치할 때만 득점하는 exact match(EM)로 이루어진다. 부분 점수가 없다는 채점 방식은 설계의 방향을 크게 제약한다. 제공된 데이터는 학습 9,534샘플, 테스트 819샘플이며, 학습 데이터의 15.5%에 해당하는 1,478샘플은 시간 순서를 정할 수 없는 것으로 표시되어 항등순열이 타깃으로 부여된다.

프레임 순서의 판별과 복원은 시각 자기지도 학습의 대리 과제로 오래 다루어져 왔다(순서 검증[4], 시퀀스 정렬[5], 시간 방향 판별[6]). 본 과제는 순서 복원 자체가 목표이고 캡션이라는 언어 조건이 함께 주어지며, 동적 해상도를 지원하는 시각-언어 모델[11]을 통해 두 양식을 하나의 시퀀스로 다룬다는 점에서 조건이 다르다.

시각-언어 모델에 순열을 텍스트로 생성시키는 직관적인 베이스라인은 두 가지 한계를 갖는다. 첫째, 미세조정 이전의 zero-shot 성능은 사실상 무작위 수준($1/24 \approx 0.042$)에 머문다. 둘째, 자유 생성은 형식 이탈과 탐색 오차를 그대로 노출하며, 각 후보에 대한 모델의 상대적 선호를 관측할 수단을 제공하지 않는다. 후보 공간이 24개에 불과하다는 사실은 이 문제를 생성이 아닌 판별 문제로 재정식화할 여지를 준다. 각 순열 π에 대응하는 답 토큰열 y_π의 조건부 로그우도 $s(\pi)$를 전수 계산하고 최대값을 취하면, 형식 이탈이 원천적으로 제거되는 동시에 1위와 2위의 점수차라는 결정 마진을 진단 신호로 확보할 수 있다.

전체 파이프라인은 다음과 같이 요약된다.

캡션 + 셔플 프레임 4장 $\rightarrow$ QLoRA 미세조정 시각-언어 모델 $\rightarrow$ 24개 후보 순열 전수 채점 $s(\pi)$ $\rightarrow$ 제시순서 2배열 점수 평균 $\bar{s}(\pi)$ $\rightarrow$ $\hat{\pi} = \arg\max_\pi \bar{s}(\pi)$

본 보고서의 기여는 다음 세 가지이다. 셋은 독립된 기법의 나열이 아니라, exact match라는 0/1 지표에 대한 위험 최소화라는 하나의 원리를 결정 규칙, 불변성, 표현의 세 층위에서 구현한 것이다.

1. **유한 순열 공간에서의 exact-risk 최소화 재정식화.** 과제를 순열의 텍스트 생성이 아니라 24개 후보 위의 위험 최소화 문제로 재정식화한다. 각 후보의 조건부 우도를 전수 채점하고 최대값을 선택하는 이 절차는 0/1 손실 하의 최소 베이즈 위험 결정과 일치하며[3], 후보 집합이 전수 열거된 조건에서 주어진 모델 분포에 대해 최적이다. 요점은 우도 채점 기법 자체가 아니라, exact-match 목적과 유한 후보 공간을 결정 이론으로 연결한 재정식화에 있다.

2. **순열 불변성 기반 증강 및 추론 설계.** 프레임의 제시 순서는 정답과 무관한 명목 변수이지만, 모델이 이를 지름길 단서로 삼을 위험이 있다[7]. 이를 차단하기 위해 학습 단계에서 제시 순서를 재배열한 4배 증강을 적용하고, 추론 단계에서는 서로 다른 두 제시 배열의 점수를 평균하여 배열 의존성을 상쇄한다[8].

3. **오류 분석에 근거한 해상도 보존 학습.** 홀드아웃 953샘플의 24순열 점수를 전량 분석하여 오답이 결정 경계 근방의 저마진 구간에 밀집함을 확인하고, 이를 근거로 원본 640×360 해상도를 보존한 학습을 수행하였다. 추론 해상도만의 변경은 효과가 없었던 반면 학습 단계의 보존은 리더보드를 개선하였다.

최종 구성은 Public 리더보드 0.93368(15위), Private 리더보드 0.92682(7위)로 마감하였다. 학습은 A100 1장에서 약 9시간이 소요되었고 산출물은 1.8 GB 어댑터이며, 외부 데이터와 외부 API는 사용하지 않았다. 단일 RTX 3090(24 GB) 환경에서의 재현은 별도로 검증하였다.

2 방법론

본 절은 캡션 1문장과 셔플된 프레임 4장으로부터 원래의 시간 순서를 복원하는 시스템의 구성 요소를 기술한다.

2.1 기반 모델과 저계수 적응

모델 규모의 상한은 추론 환경의 자원 제약에 의해 결정된다. 최종 제출물은 RTX 3090 24 GB 단일 장치에서 24시간 이내에 테스트 819샘플 전체를 처리할 수 있어야 한다. 전수 채점 디코딩은 샘플당 48개 시퀀스를 배치로 채점하므로 활성화 메모리와 처리량이 함께 상한을 규정한다. 4-bit NF4 양자화[2]를 전제로 할 때 밀집(dense) 구조 기준 27B급이 이 조건을 만족하는 최대 규모였다(실측은 3.8절). 기반 모델 후보 비교와 신형 교체 시도는 3.2절에서 다룬다.

적응은 4-bit로 양자화된 기반 모델 위에 저계수 어댑터[1]를 학습하는 QLoRA[2] 방식을 따른다. 어댑터는 언어부의 attention 및 MLP 투영에만 부착하고 비전 인코더는 동결하였으며, 학습되는 파라미터는 전체의 1.68%에 해당한다. 랭크 $r = 64$, 스케일 $\alpha = 64$, dropout 0.05를 사용하고 학습률은 2×10^{-4} cosine 스케줄에 warmup 3%를 적용하였다. 상세 구성은 표 1에 정리하였다.

2.2 전수 채점 디코딩

정답 공간은 4프레임의 순열 24개로 유한하다. 따라서 자유 생성으로 순열을 산출할 필요가 없으며, 24개 후보를 모두 강제 디코딩하여 조건부 로그우도를 비교하는 절차가 타당하다. 캡션을 c, 제시배열 σ로 배치된 프레임을 X_σ, 순열 π에 대응하는 정답 토큰열을 y_π라 할 때 후보 점수는

$$s(\pi) \;=\; \sum_{t} \log p_\theta(y_{\pi,t} \,|\, y_{\pi,<t},\, c,\, X_\sigma) \tag{1}$$

이며, 예측은 $\hat{\pi} = \arg\max_\pi s(\pi)$로 정한다. 24개 후보의 토큰 길이가 동일하므로 길이 정규화는 불필요하다.

이 결정 규칙은 최소 베이즈 위험 디코딩[3]의 특수한 경우이다. 평가 손실이 시퀀스 단위 0/1 손실일 때 베이즈 위험을 최소화하는 결정은 사후확률이 최대인 후보를 선택하는 것과 일치하며, 후보 집합이 24개로 전수 열거 가능하므로 근사 없이 정확한 해를 얻는다. 즉 학습 시의 목적(정답 토큰열의 우도 최대화)과 추론 시의 결정 규칙, 그리고 대회의 채점 지표가 동일한 목적 위에 정렬된다.

이 방식에서는 형식 위반 출력이 원천적으로 배제되고 추론이 결정론적이며, 모든 후보의 점수가 함께 산출되어 1–2위 점수차를 결정 마진이라는 진단 신호로 활용할 수 있다(3.6절).

2.3 순열 불변 학습을 위한 재배열 증강

정답은 프레임의 내용에만 의존해야 하고 프레임이 제시된 순서에는 의존하지 않아야 한다. 그러나 원자료는 각 샘플에 대해 하나의 제시배열만 제공하므로, 모델이 내용 대신 제시 위치에 결부된 규칙성을 학습할 여지가 남는다. 이러한 지름길 학습[7]을 차단하기 위해 각 학습 샘플의 프레임 제시순서를 서로 다른 배열로 4회 재배열하고 그에 맞추어 타깃 순열을 재계산하여, 9,534샘플로부터 38,136개의 학습 예제를 구성하였다. 이로써 위치에 기반한 규칙은 학습 신호로서 상쇄되고 내용에 기반한 순서 판단만 남는다.

증강의 종류는 의도적으로 제한하였다. 좌우 반전이나 시간 축 반전 같은 기하·시간 변형은 일반적인 영상 인식에서는 무해하지만, 본 과제에서는 판단 근거 자체를 훼손한다. 시간 방향 단서는 운동의 방향성과 비가역적 사건 진행에 실려 있으므로[6], 이러한 변형은 입력을 바꾸면서 라벨의 의미를 무효화한다. 색상·잡음 계열 변형 역시 프레임 간 미세한 상태 변화를 흐릴 위험이 있어 배제하였다.

정렬 불가로 표기된 샘플 1,478건(15.5%)은 모든 재배열에서 타깃을 항등순열로 유지한다. 재배열 증강은 정렬 가능한 샘플에서 나머지 순열의 노출을 균등화하여 항등순열로의 기본값 편향을 억제하는 역할도 겸한다.

2.4 순열 자기일관성 시험 시간 증강

학습 단계에서 제시순서 불변성을 유도하더라도 잔여 편향은 남는다. 리스트 단위 순위 매기기에서 입력 제시순서에 따른 편향을 서로 다른 배열의 출력을 집계하여 상쇄하는 순열 자기일관성[8]의 원리를 추론 단계에 적용하였다. 구체적으로 동일 샘플에 대해 K개의 제시배열 $\sigma_1, \ldots, \sigma_K$를 만들고, 각 배열에서 원래 프레임 인덱스 기준으로 정렬한 후보 점수를 평균한다.

$$\bar{s}(\pi) \;=\; \frac{1}{K} \sum_{k=1}^{K} s^{(\sigma_k)}(\pi), \qquad \hat{\pi} = \arg\max_{\pi} \bar{s}(\pi) \tag{2}$$

전수 채점 덕분에 순위가 아니라 점수 자체를 평균하는 집계가 가능하다. 이는 복수 관측의 평균으로 개별 관측의 분산을 줄이는 자기일관성[9] 및 시험 시간 증강 집계[12]와 같은 구조로, 결정 마진이 작은 샘플의 배열 의존적 요동을 완화한다. 최종 시스템은 $K = 2$를 사용하며, 집계 대상이 동일 모델의 관측이므로 앙상블이 아니라 단일 모델의 입력 측 증강이다.

2.5 원본 해상도 보존 학습

동적 해상도를 지원하는 시각-언어 모델[11]에서 입력 픽셀 예산은 시각 토큰 수를 통해 연산량과 표현 정밀도를 동시에 결정하는 하이퍼파라미터이다. 본 과제의 원본 프레임은 640×360이며, 관행적인 축소 설정은 이를 리사이즈하여 프레임 간 미세한 차이를 소실시킨다. 오류 분석(3.6절)에서 오답이 판별 경계 근방에 밀집함이 확인되었으므로, `max_pixels`를 307,200으로 설정하여 원본 해상도를 무손실로 보존한 상태에서 학습과 추론을 모두 수행하였다. 개입 시점에 따른 대비는 3.7절에서 제시한다.

3 실험

3.1 실험 설정

과제 정의와 평가 지표는 1절을 따른다. 학습 집합은 9,534샘플, 테스트 집합은 819샘플이다. 학습 집합 가운데 1,478건(15.5%)은 정렬이 불가능한 것으로 표기되어 있으며 항등 순열이 타깃으로 부여되어 있다. 학습 캡션의 약 87%는 *then*, *before*, *after* 등의 시간 접속사를 포함하고 있어, 언어 측 단서는 형식상 충분히 제공된다.

학습 구성은 표 1에 정리하였다.

검증은 고정 시드 10% 홀드아웃(953샘플)으로 수행하였다. 제출이 하루 2회로 제한되므로 채택과 기각의 대부분을 오프라인에서 결정하고 리더보드는 확정된 후보의 확인 용도로만 사용하였다. 동일한 기반 모델 안에서는 홀드아웃 EM의 변화량과 리더보드 점수의 변화량이 거의 1:1로 전이됨을 반복 확인하였다. 다만 기반 모델을 교체하는 경우에는 이 전이가 성립하지 않았다(3.2절). 레시피를 동결한 뒤에는 홀드아웃 953샘플을 학습에 환원하였다.

Table 1: 최종 학습 구성

항목	설정
기반 모델	Qwen3.5-27B (4-bit NF4)
어댑터	LoRA $r = 64$, $\alpha = 64$, dropout 0.05
적용 범위	언어부 attention + MLP (비전 인코더 동결)
학습 파라미터	전체의 1.68%
학습률	2×10^{-4}, cosine, warmup 3%
학습량	1에폭, 2,384스텝, 유효 배치 16
난수 시드	3407
입력 해상도	`max_pixels` 307,200 (640×360 무손실)
증강	제시 순서 재배열 ×4 (38,136예제)
자원	A100 80GB 1장, 약 9시간

Table 2: 초기 기반 모델 비교. 동일 QLoRA 조건·전수 채점 디코딩, 초기 홀드아웃 96샘플 기준. 미세조정 이전 zero-shot은 전 모델이 무작위 수준(≈0.042)이다.

모델	EM
Qwen3-VL-30B-A3B	0.750
Qwen3-VL-8B	0.698
GLM-4.1V-9B	0.688
InternVL3.5-8B	0.656

3.2 기반 모델 선정

초기 단계에서 동일한 QLoRA 조건과 동일한 프롬프트 아래 네 종의 공개 시각-언어 모델을 비교하였다. 홀드아웃 96샘플에 대한 EM은 표 2와 같다. 모델 간 격차는 최대 0.094로, 같은 시점에 시도한 어떤 디코딩 개선이나 정칙화 개입보다 컸다. 이에 따라 이후 자원은 디코딩 개선보다 기반 모델의 규모 확장에 우선 배분하였다. 소표본 홀드아웃에서 최고였던 혼합 전문가 구조(30B-A3B)의 우위는 리더보드로 그대로 전이되지 않았고, 밀집 구조에서 규모를 확장한 Qwen3.5-27B 구성이 가장 높은 점수를 기록하여 이를 채택하였다.

규모 확장의 유효성은 최신 모델로의 교체로 이어지지 않았다. 동일 레시피를 신형 Qwen3.6-27B에 적용한 결과 0.89528로, 동일 조건의 Qwen3.5-27B(0.917) 대비 0.022 낮았고, 혼합 전문가 구조인 Qwen3.6-35B-A3B는 동일 학습률과 배치에서 발산하여 유효한 체크포인트를 얻지 못하였다. 이에 따라 기반 모델 교체는 레시피 전체의 재조율이 필요한 개입으로 보고 대회 일정 내에서는 추진하지 않았다.

3.3 주요 결과

리더보드 점수의 진행은 표 3에 정리하였다. 출발점은 홀드아웃 953샘플을 제외한 8,581샘플로 학습한 어댑터에 24개 순열 후보 전수 채점을 결합한 구성으로 0.917이었다. 여기에 프레임 제시 순서를 바꾼 두 배열의 점수를 정렬 후 평균하는 순열 자기일관성 시험 시간 증강을 더하여 0.919를 얻었다. 이후 레시피를 동결하고 홀드아웃을 학습에 환원하여 0.92844로 0.0094 상승하였으며, 마지막으로 학습 단계부터 원본 해상도를 보존하여 0.93368에 도달하였다. 이 구성이 최종 제출이며 0.93368은 Public 리더보드 점수이다. Private 리더보드에서는 0.92682로 7위에 해당한다.

증분의 크기는 개입의 성격에 따라 구분된다. 데이터를 늘리는 개입과 입력 정보를 보존하는 개입은 각각 0.0094와 0.00524의 이득을 남긴 반면, 디코딩 단계의 집계는 0.002에 그쳤고 $K \geq 3$에서는 포화하여 추가 이득이 없었다. 이득의 대부분은 모델이 관측하는 입력과 관측 횟수에서 발생하였고, 출력 후처리 계열 개입의 개선 여지는 제한적이었다.

3.4 구성 요소 제거 실험

표 4은 실제 제출 또는 홀드아웃 평가로 확인한 개입들을 정리한 것이다. 대부분은 개선이 없거나 성능을 떨어뜨렸으며, 부정적 결과 자체가 제약 요인의 위치를 좁히는 근거로 사용되었다.

채점 목적함수도 비교하였다. 캡션의 우도를 평가하는 생성 방향은 홀드아웃 0.39로, 순열 인덱스를 직접 채점하는 방향(0.51)에 뒤졌다. 긴 캡션을 채점할수록 신호가 언어적 유창성에 흡수되는 것으로 해석된다.

생성 우도가 아닌 은닉 표현 위의 판별기를 쓰는 방향은 세 가지 형태로 구현하여 모두 검증하였다. 첫째, 마지막 프롬프트 토큰의 표현에서 24개 순열을 직접 분류하는 헤드는 생성 채점을 넘어서지 못하였다. 둘째, 프레임별 표현으로 6개 쌍의 선후를 판별하는 헤드는 쌍별 정확도 0.62에서 정체하였다. 인과 어텐션 하에서 k번째 프레임의 표현은 이후 프레임을 관측하지 못하므로 양방향 비교가 구조적으로 불가능하기 때문이다. 셋째, 이 제약을 우회하여 모든 프레임을 관측한 마지막 프롬프트 토큰의 표현에서 6개 쌍의 선후를 예측하는 생성-판별 멀티태스크를 구성하였다. 학습은 성공하여 학습 집합 쌍별 정확도 0.93에 도달하였으나, 홀드아웃에서 판별 경로 단독 EM은 0.379로 같은 모델의 생성 채점 0.600에 크게 못 미쳤다. 두 점수의 선형 결합 또한 모든 가중치에서 이득이 없었고 최적 가중치는 0이었다. 판별 헤드가 생성 채점이 담지 못한 정보를 갖고 있지 않음을 의미하며, 병목이 판독기 설계가 아니라 입력 표현의 분해능에 있음을 시사한다.

3.5 과적합 회피

학습량과 어댑터 스케일을 상향한 두 개입은 학습 손실을 낮추면서 리더보드를 떨어뜨렸다. 2에폭 학습은 0.871로 0.046 하락하였고, LoRA 스케일을 $\alpha = 2r$로 올린 구성은 학습 손실이 기준보다 낮았음에도 리더보드는 0.90401로 0.024 하락하였다. 후술할 라벨 감사에서 학습 라벨의 약 5%가 노이즈로 의심된다는 점을 고려하면, 이 손실 감소의 상당 부분은 일반화 가능한 규칙이 아니라 노이즈 라벨의 암기에 배분되었다고 보는 것이 합리적이다[10]. 실제로

Table 3: Public 리더보드 점수 진행. 증분은 직전 행 대비이다.

구성	LB	증분
SFT(8,581) + 전수 채점	0.917	—
+ 순열 자기일관성 TTA ($K = 2$)	0.919	+0.002
+ 전체 데이터 재학습	0.92844	+0.0094
+ 해상도 보존 학습	0.93368	+0.00524

Table 4: 구성 요소 제거 및 대안 개입 결과. 괄호는 해당 시점의 기준 구성 대비 변화량이다.

개입	동기	측정	결과
2에폭 학습	적합도 상향	LB	0.871 (−0.046)
LoRA $\alpha = 2r$	적응 폭 확대	LB	0.90401 (−0.024)
라벨 노이즈 정제 + 커리큘럼	노이즈 억제	LB	−0.029
난수 시드 1234 재학습	실행 분산 추정	LB	0.92495 (−0.0035)
추론 해상도만 640	저비용 해상도 개입	홀드아웃 EM	±0.0000
TTA $K \geq 3$	집계 확대	LB	포화(추가 이득 없음)
비전 인코더 해동(448)	시각 표현 적응	홀드아웃 EM	이득 없음
가중치 평균화 및 어댑터 병합	실행 분산 저감	홀드아웃 EM	이득 없음
선호 학습(SimPO/DPO/IPO)	순위 목적 직접 최적화	홀드아웃 EM	이득 없음
listwise 교차엔트로피	순위 목적 직접 최적화	홀드아웃 EM	이득 없음
캡션 조건부 채점 목적	역방향 우도 활용	홀드아웃 EM	0.39 (답 토큰 0.51)

라벨 노이즈를 정제하고 커리큘럼을 적용한 구성은 0.029 하락하였는데, 이는 노이즈 표본의 제거가 동시에 항등 순열 분포에 대한 정보를 함께 제거하기 때문으로 해석된다. 최종 구성은 1에폭, $\alpha = r$, 단일 시드를 유지하는 보수적 설정을 택하였다. 공개 리더보드 표본 과적합의 위험을 줄이는 방향을 우선한 결정이다. 실제로 본 구성은 Public 리더보드 15위에서 Private 리더보드 7위로 상승하였는데, 이는 상위권의 상당수가 Public 분할에 과적합된 반면 본 구성의 성능은 분할 간에 비교적 안정적으로 유지되었음을 보여준다. 시드 1234로 재학습한 구성이 0.92495로 0.0035 낮았다는 점은 이 규모의 차이가 개입의 효과가 아니라 학습 잡음일 수 있음을 보여주므로, 0.005 미만의 증분은 단일 실행만으로 채택하지 않았다.

3.6 오류 분석

중간 체크포인트(홀드아웃 953샘플, EM 0.600)에 대해 24개 순열의 점수를 전량 저장하고 오답 381건을 분석하였다. 예측과 정답 사이의 인접 전위 거리(Kendall tau 거리) 분포는 1회 스왑 130건(34%), 2회 69건, 3회 78건, 4회 51건, 5회 31건, 6회 22건(6%)이다. 완전 역순을 포함하는 최대 거리 오류가 6%에 불과하다는 사실은 모델이 시간의 방향 자체를 반대로 이해하는 실패는 거의 일으키지 않음을 뜻한다[6].

점수 분포는 이 국소성을 다른 각도에서 확인해 준다. 오답 케이스에서 1위와 2위 후보의 점수 차이(마진) 중앙값은 0.027로, 정답 케이스의 0.109에 비해 약 4분의 1 수준이다. 오답의 40%는 마진이 0.02 미만이었다. 인접 스왑 오답 130건 가운데 정답이 2위인 경우가 57건(44%)이라는 점까지 함께 보면, 상당수의 오류는 모델이 정답을 후보군에서 배제한 결과가 아니라 근소한 차이로 순위를 뒤바꾼 결과이다.

그러나 국소성이 오류 전체를 설명하지는 않는다. 인접 스왑 오답 130건을 정답보다 높은 점수를 받은 순열들의 구성으로 분해하면, 정답 위에 놓인 후보가 전부 인접 스왑인 순수 국소 오류는 60건(46%)이고, 인접하지 않은 순열까지 정답보다 높게 평가된 광역 실패가 70건(54%)으로 오히려 많다. 즉 표면적으로 한 번의 인접 교환으로 보이는 오류의 절반 이상은 인접 프레임 쌍의 미세한 판별 실패가 아니라 순서 구조 전반에 대한 확신 부족에서 비롯된다. 인접 쌍 판별만을 보강하는 접근으로 회수할 수 있는 오류가 제한적임을 의미하며, 앞 소절의 쌍별 판별 헤드가 실패한 결과와 일관된다.

시간 접속사를 포함한 캡션의 비율은 오답군에서 83%로 전체 87%와 차이가 없었다. 실패는 언어 단서의 유무가 아니라 이를 시각 증거에 대응시키는 단계에서 발생한다.

마지막으로 라벨 노이즈를 감사하였다. 충분히 학습된 모델이 주어진 정답을 24개 후보 중 하위권으로 평가하는 샘플이 약 5% 존재한다. 이 가운데 일부는 모델의 실패이겠으나 일부는 라벨 자체의 문제로 보이며, 테스트 집합에도 유사한 비율의 노이즈가 있다면 달성 가능한 정확도의 상한 자체가 제한된다. 이는 3.5절의 관찰과 부합한다[10].

종합하면 오류의 다수는 결정 경계 근방의 저마진 샘플에 집중되며, 그 저마진은 판독기 설계가 아니라 시각 입력 표현의 분해능에서 비롯된다는 가설에 도달한다. 마진이 작다는 것은 표현이 이미 정답 경계에 근접해 있다는 뜻이므로, 결정 규칙을 더 손보기보다 표현의 충실도를 높이는 개입이 이 표본들을 회복시킬 가능성이 가장 크다. 다음 소절의 해상도 실험이 이 예측을 직접 검증한다.

3.7 해상도 개입 시점 분석

원본 640×360 프레임을 448 기준으로 축소하여 학습한 구성과 원본을 보존한 구성을 비교하기에 앞서, 해상도가 추론 시점에만 개입할 때의 효과를 짝지어 측정하였다. 448로 학습된 동일 어댑터에 추론 해상도만 640으로 올리면 홀드아웃 EM 변화는 정확히 0.0000이었다. 예측이 바뀐 샘플은 88건이었고 그중 오답에서 정답으로 바뀐 것이 23건, 정답에서 오답으로 바뀐 것이 23건으로 상쇄되었으며, 나머지 42건은 오답 사이의 이동이었다. 이 88건의 결정 마진 중앙값은 0.0156으로, 예측이 유지된 안정 샘플의 0.0859에 비해 약 6분의 1이다. 추론 해상도 상향은 결정 경계 근방의 샘플만 무작위로 뒤집었을 뿐이다.

반면 학습 단계부터 원본 해상도를 유지하면 리더보드가 0.00524 상승하였다. 두 구성의 학습 손실 곡선은 사실상 구분되지 않았다. 평균 학습 손실은 640에서 0.140, 448에서 0.144로 차이가 미미하였다. 총 손실이 거의 같은데 EM

Table 5: 제출본과 재현 실행이 상이한 4건의 명세. 결정 마진은 1–2위 후보의 길이정규화 로그우도 차이이다.

Id	제출본	재현 실행	결정 마진
XGuH3q	[1, 3, 2, 4]	[2, 3, 1, 4]	0.0107
9uf8Mn	[3, 4, 1, 2]	[3, 4, 2, 1]	0.0011
HPk5qW	[1, 3, 2, 4]	[1, 4, 2, 3]	0.0078
LRHWnG	[2, 1, 4, 3]	[3, 1, 4, 2]	0.0041

만 개선된 것은, 해상도 보존이 전반적 적합도가 아니라 결정 경계 근방의 판별 정보를 개선하였음을 뜻한다.

두 실험은 개입 시점이 결과를 좌우함을 보여준다. 동일한 픽셀 정보라도 학습 과정에서 주어져야 판별에 기여하는 표현으로 편입되며, 이는 시각 표현의 충실도가 제약 요인이라는 가설과 정합적이다.

3.8 자원 및 재현성

재현성 요구를 충족하기 위해 RTX 3090 24 GB, CUDA 12.4 환경에서 동작하는 별도 추론 경로를 구현하였다. 학습에 사용한 가속 라이브러리에 의존하지 않고 순정 transformers와 peft만으로 구성하였으며, 완전 오프라인 로딩을 포함한 테스트 819건 전량 재실행에서 815건(99.5%)이 제출본과 일치하였다. 상이한 4건의 명세는 표 5와 같다. 실행 스택 간 후보 점수의 수치 섭동을 직접 측정한 결과 후보당 차이는 중앙값 0.004, 최대 0.018로, 상이 4건의 결정 마진(0.001–0.011)과 같은 자릿수인 반면 안정 샘플의 마진(중앙값 0.086)보다 한 자릿수 이상 작았다. 즉 불일치는 결정 마진이 커널 수치 섭동보다 작은 극소수 표본(약 0.5%)에 국한되는 부동소수 수준의 현상이며, 점수 영향은 ±0.005 이내이다. `micro_batch` 4 기준 최대 메모리 점유는 21.73 GB, 테스트 819샘플 전체 추론 시간은 3090 환산 약 15.8시간으로 24시간 제한 안에 들어오며, 메모리 부족 시 배치를 4에서 2, 1로 자동 강등한다.

4 결론

본 보고서는 셔플된 프레임 4장의 시간 순서 복원 과제를, 4-bit NF4 양자화 위의 저계수 적응[2, 1]으로 미세조정한 단일 시각-언어 모델과 24개 순열 전수 채점 디코더의 결합으로 다루었다. 제시 순서 재배열 증강과 원본 해상도 보존 학습, 두 배열의 순열 자기일관성 집계를 더한 최종 구성은 Public 리더보드 0.93368(15위), Private 리더보드 0.92682(7위)를 기록하였으며, Public 대비 Private에서의 순위 상승은 과적합을 경계한 설계의 사후 검증에 해당한다. 앙상블 없이 단일 모델과 1.8 GB 어댑터로 얻은 결과이고, 추론은 단일 RTX 3090에서 대회의 자원 제약을 충족한다.

방법론의 핵심은 채점 지표에 정합적인 결정 규칙[3]과 순열 불변성[8, 7]이라는 두 원리로 학습·증강·추론을 일관되게 정렬한 데 있다. 아울러 대부분의 채택과 기각을 오프라인 검증으로 결정하고, 기각된 시도들을 수치와 함께 기록하였다(표 4). 본 사례가 시사하는 일반 원리는 다음과 같다. 출력 공간이 작고 유한하며 평가가 전부 아니면 전무인 과제에서는, 자유 생성을 후보 전량의 우도 채점으로 대체하는 것이 지표에 정합적인 결정 규칙을 제공한다. 그 이후의 개선은 디코더가 아니라 결정 경계 근방의 마진을 키우는 입력과 표현 측의 개입에서 나온다.

한계는 다음과 같다. 각 프레임을 독립 이미지로 인코딩하는 현 구조는 화소 수준 대응에서 드러나는 미세 운동 단서를 표현 단계에서 활용하지 못하며, 선두권과의 격차는 이 제약에서 비롯한 것으로 추정된다. 판별 헤드와 오류 분석의 결과도 남은 실패가 비전 인코더가 구분하지 못한 인접 프레임 쌍에 집중됨을 가리킨다. 라벨 노이즈(약 5%)가 테스트 분할에도 존재한다면 달성 가능한 정확도에는 상한이 걸리고[10], 전수 채점은 후보 수가 계승적으로 늘어나는 5프레임 이상의 설정으로 그대로 확장되지 않는다. 향후 과제는 프레임 간 대응을 표현 수준에서 다루는 시간 인식 백본의 도입과, 전수 채점을 대체할 후보 가지치기 기반 근사 디코딩의 설계이다.

References

[1] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *NeurIPS*, 2023.

[3] S. Kumar and W. Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *HLT-NAACL*, 2004.

[4] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*, 2016.

[5] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017.

[6] D. Wei, J. Lim, A. Zisserman, and W. T. Freeman. Learning and using the arrow of time. In *CVPR*, 2018.

[7] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 2020.

[8] R. Tang, C. Zhang, X. Ma, J. Lin, and F. Ture. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *NAACL*, 2024.

[9] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[10] D. Arpit et al. A closer look at memorization in deep networks. In *ICML*, 2017.

[11] P. Wang et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024.

[12] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag. Better aggregation in test-time augmentation. In *ICCV*, 2021.