구조화 순열 예측과 저신뢰 잔차 보정을 이용한 멀티모달 장면 순서 복원

팀 김수민딥기시쁠

김수민 김형진 성민수 조희재

kimsm315@snu.ac.kr yolo04250@snu.ac.kr minsnu1234@snu.ac.kr choheejae@snu.ac.kr

초록

본 과제는 하나의 캡션과 무작위 순서로 제시된 네 프레임으로부터 원래의 시간 순서를 복원하는 멀티모달 순열 예측 문제다. 출력이 24개 유효 순열로 제한된다는 점을 이용해 자유 텍스트 생성 대신 모든 후보를 직접 점수화하는 정식화를 택했고, Qwen3.5-27B의 전역 표현과 프레임별 표현 위에서 direct 분류, frame rank, pairwise 선후관계를 동시에 학습하였다. 이후 기반 모델 전체를 동결하고 같은 head의 내부 증거만 읽는 파라미터 6,641개의 잔차 모듈을 추가해, margin 게이트가 저신뢰 예측만 보정하도록 하였다. 본 보고서의 방법은 오류의 구조를 먼저 측정하고 그 측정값이 출력 공간·손실 가중치·보정 대상을 결정하도록 구성한 것이 특징이며, 설계가 의도한 지점에서만 작동했는지를 보정 전후 대조로 검증하였다. 최종 모델은 고정 검증 954건에서 Exact Match 52.52%를 기록했고, 예선 test에서 public 0.93542, private 0.91463, 도메인이 다른 외부 검증 test의 공개 구간(50건)에서 0.78을 기록하였다.

1 서론

각 샘플은 캡션 하나와 입력 순서가 섞인 네 프레임으로 구성되며, 네 프레임의 시간 rank를 정확히 복원해야 Exact Match로 인정된다. 답이 24개 순열의 폐집합이므로 자유 생성은 파싱 실패와 무효 출력의 위험만 더한다. 우리는 24개 후보를 직접 점수화하는 분류 정식화를 택했고, 실제로 동일 규모의 생성형 판독(verbalizer SFT)과 대조 실험한 결과 검증 EM 49.79% 대 49.58%로 이점이 없음을 확인한 뒤 이 선택을 고정하였다.

본 보고서의 기여는 다음 네 가지다. (1) 검증 오답을 confidence, identity 편향, 캡션 길이, 순열 오류 유형의 네 축으로 분해해 약점을 특정하고, 각 약점을 보정 모듈의 특징·게이트·손실 항에 일대일로 대응시킨 진단·설계·검증 사슬. (2) direct, rank, pairwise 세 경로를 24개 후보 순열의 공간으로 사영해 가산 결합하는 구조화 head. (3) 기반 모델을 완전히 동결한 채 내부 증거만 다시 읽는 저신뢰 잔차 보정으로, 대회의 단일 모델·단일 checkpoint 규정을 준수. (4) 보정이 의도한 구간에서만 작동했는지의 사후 정량 검증으로, 예측이 바뀐 169건 중 168건(99.4%)이 게이트 개방 구간에서 발생했고 차단 구간 335건은 한 건도 바뀌지 않았음을 확인.

2 데이터 전처리 기법

시각적으로 거의 같은 장면이 학습과 검증 양쪽에 들어가면, 검증이 순서 추론 능력이 아니라 장면 재인식을 측정하게 된다. 이를 막기 위해 제공 train 9,535건 내부에서만 지각적 중복 그래프를 만들었다. 각 프레임의 밝기 표준편차가 5 이상일 때만 64비트 pHash를 계산하고, Hamming 거리 6 이하를 프레임 수준 근접일치로 정의한 뒤, 같은 프레임을 재사용하지 않는 최대 이분 매칭으로 두 행 사이의 서로 다른 일치 프레임 수를 세었다. 세 프레임 이상 일치한 행 쌍을 Union–Find로 연결해 9,261개 그룹(강한 연결 305쌍, 최대 그룹 8행)을 얻었다. 검증셋은 그룹을 쪼개지 않은 채 크기 오차와 24개 정답 class 빈도 오차의 가중합을 최소화하도록 선택했고(seed 42, 무작위 greedy 128회 재시작), 선택된 검증셋과 두 프레임 이상 직접 일치하는 비검증 행은 해당 그룹 전체를 개발 학습에서 제외(purge)하였다. purge된 2,974건은 미리 정한 비율이 아니라 이 경계 규칙의 결과이며, 레시피 확정 후 최종 학습에는 다시 포함된다. 후보 쌍 34,006개를 전수 검사한 결과 최종 분할에서 학습과 검증 사이에 남은 직접 근접일치 쌍은 0건이다.

Table 1: 고정 개발 분할. 세 부분 어디에서도 근접중복 그룹이 경계를 넘지 않는다(교차 그룹 0개).

역할	행 수	그룹 수	개발 단계 역할
Development train	5,607	5,496	gradient 학습
Validation	954	931	레시피 고정 평가
Purged	2,974	2,834	경계 누출 보호를 위해 보류

Table 1의 그룹 수 합(5,496+931+2,834=9,261)이 전체 그룹 수와 일치하므로, 어떤 그룹도 두 부분에 걸쳐 쪼개지지 않았음이 표 안에서 검산된다. 한 가지 사후 관찰로, 분할의 층화 대상은 누출 차단과 class 빈도였고 캡션 길이는 아니었다. 그 결과 짧은 캡션(24단어 이하) 비율이 train 60.4%, validation 45.4%, purged 9.3%로 갈렸다. 이는 6절에서 논의할 검증셋 대표성 한계의 원인이 된다. test 데이터의 내용·정답·분포는 분할과 설정 선택에 사용하지 않았다.

3 모델 구조와 설계

설계는 24개 class에 대한 direct 분류에서 출발해 절대 rank와 상대 pair 감독을 더하고, causal 프레임 표현에 context fusion을 적용한 뒤, 마지막으로 동결된 head 위에 저신뢰 잔차를 부착하는 순서로 발전했다. 성능 기법을 병렬로 나열한 것이 아니라 앞 단계의 오류 분석이 다음 설계의 입력이 되는 점진 구조다(Figure 1).

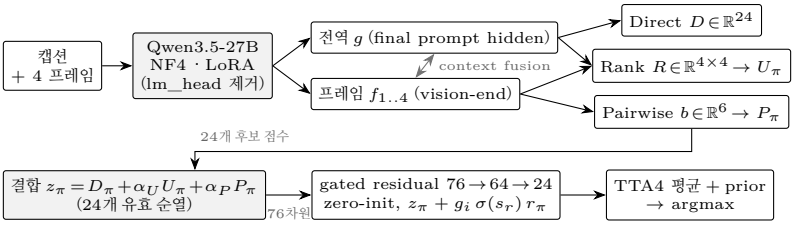

Figure 1: 전체 계산 경로. 백본의 전역·프레임 표현이 세 branch(direct·rank·pairwise)로 갈라져 24개 유효 순열 공간에서 결합되고, 동결된 head의 내부 증거를 다시 읽는 gated residual이 저신뢰 구간만 보정한다. 하나의 base·adapter·augmented head만 사용하며 서로 다른 모델·seed·epoch의 예측을 결합하지 않는다.

3.1 Qwen3.5-27B 백본 선택

Qwen3.5-27B는 비전 타워(ViT 27층)와 하이브리드 어텐션 텍스트 타워(64층 중 full attention 16층, 나머지는 선형 어텐션)를 갖는 dense 멀티모달 모델이다. 백본은 세 단계를 거쳐 확정했다. 8B급 분류 head 계열은 우리 팀을 포함해 여러 팀이 public 0.895 부근에 수렴해 접근법의 천장으로 판단했고, Qwen3-VL-32B로 규모를 키워 0.911을 얻었으나 추론 peak 예약 메모리 31.4 GiB로 검증 환경(24 GB) 초과 위험이 있었다. 이에 레시피 전체를 고정한 채 백본만 교체하는 one-factor 실험으로 Qwen3.5-27B를 시험했고, 파라미터가 오히려 20% 줄었음에도 public이 0.92844로 상승했으며 추론 peak가 17.2 GiB로 내려와 성능과 배포 제약을 동시에 해결했다(Table 2). 성능 향상의 주 원인이 규모가 아니라 백본 세대임을 보여 준다. 분류 경로만 사용하므로 어휘 투영 lm_head(약 1.27B 파라미터)는 메모리에서 제거해 양자화 준비 과정의 불필요한 상승을 막았다.

Table 2: 백본 one-factor 궤적. 레시피(LoRA·손실·TTA·prior)는 동일하게 고정하였다.

백본	예선 public	추론 peak VRAM	비고
Qwen3-VL-8B	0.895 부근	—	여러 팀 공통 천장
Qwen3-VL-32B	0.91099	31.4 GiB	24 GB 환경 초과 위험
Qwen3.5-27B (epoch 2)	**0.92844**	**17.2 GiB**	파라미터 −20%, 세대 교체

3.2 구조화 순열 예측

모델 입력은 네 프레임과 캡션을 하나의 토큰 열로 이어 붙인 것이다. causal 구조에서는 마지막 토큰만이 전체 입력을 읽은 위치이므로 마지막 유효 prompt 토큰(padding 제외)의 hidden을 전역 특징 g로, 각 이미지의 끝을 표시하는 vision-end 토큰 네 곳의 hidden을 프레임 특징 $f_1..f_4$로 사용한다. 같은 causal 구조 때문에 앞쪽 프레임 토큰은 뒤쪽 이미지를 보지 못하므로, 각 프레임을 다음과 같이 전역 특징과 융합한다.

$$f_i' \;=\; f_i \;+\; \sigma(s_c)\cdot \mathrm{MLP}\big([\,f_i;\; g;\; f_i \odot g;\; f_i - g\,]\big).$$

단순 합 $f_i + g$는 모든 프레임에 같은 값을 더해 24개 후보 점수의 차이에서 상쇄되므로, 곱·차 상호작용 항이 공유 문맥을 프레임마다 다르게 작용시키는 역할을 한다.

세 branch가 후보 순열 π의 점수를 만들며, 각 branch는 hidden을 512차원으로 사영한 뒤 점수를 계산한다. direct branch는 g로부터 24개 class의 logits D를 직접 출력하고, rank와 pairwise branch는 융합 프레임 특징 f'로부터 계산된다. rank branch의 $R \in \mathbb{R}^{4\times 4}$는 $U_\pi = \sum_i R_{i,\pi_i}$로, pairwise branch의 6개 선후 logits b는 부호합 $P_\pi = \sum_{i<j} s_{ij}(\pi)\, b_{ij}$($\pi$가 i를 j보다 앞에 두면 $s_{ij} = +1$, 아니면 -1)로 변환된다. 최종 점수는

$$z_\pi \;=\; D_\pi + \mathrm{softplus}(s_U)\, U_\pi + \mathrm{softplus}(s_P)\, P_\pi$$

이며, 학습 종료 시 $\alpha_U = \mathrm{softplus}(s_U) = 0.485$, $\alpha_P = \mathrm{softplus}(s_P) = 0.417$로 세 경로가 모두 실질적으로 기여했다. 모든 계산이 열거된 24개 순열 위에서만 이루어지므로 argmax는 항상 전역적으로 유효한 순열이다. 예컨대 한 검증 행에서 direct만으로는 identity가 1.2 대 0.7로 우세했지만, rank·pair 점수를 결합하자 2.74 대 4.06으로 역전되어 올바른 순열이 선택되었다. 24개 class 분류에서 인접 순열은 서로 독립인 라벨이지만, rank·pair 사영은 순열 공간의 기하를 공유하기 때문이다. 순열 분류 [4]와 쌍별 선후 감독 [5]은 자기지도 영상 표현 학습에서 개별적으로 연구된 신호이며, 본 설계의 차이는 세 감독을 한 순열 공간에서 결합하고 손실 가중을 오류 분석으로 정한 데 있다.

학습 목적함수는 결합 logits의 주 교차엔트로피에 세 보조 감독을 더해 다음과 같이 구성된다.

$$\mathcal{L}_{\mathrm{base}} = \mathrm{CE}(z, y) + 0.1\,\mathrm{CE}(D, y) + 0.1\,\tfrac{1}{4}\textstyle\sum_i \mathrm{CE}(R_i, y_i) + 0.5\,\mathrm{BCE}(b, t),$$

여기서 t는 6개 쌍의 선후 정답이며, label smoothing 0.03을 CE 항에 적용했다. pairwise 가중을 기본값보다 높인 0.5로 둔 근거는 오류 분석이다. 긴 캡션 오답의 58%가 인접 두 프레임의 선후 혼동(인접 2-swap)이었으므로, 쌍별 선후 판단 하나가 그 쌍을 포함하는 12개 순열에 동시에 반영되는 pairwise 경로를 강화하였다.

3.3 진단에 기반한 저신뢰 잔차 보정

개발 모델의 고정 검증 954건(정답 494, 오답 460)을 분해해 네 가지 약점을 특정했고, 각 약점을 보정 모듈의 설계 요소에 일대일로 대응시켰다(Table 3). 핵심 관찰은 모델의 신뢰도가 이미 잘 보정되어 있다는 점이다. margin(1위−2위 점수) 구간별 정확도가 21.1%에서 100%까지 완전 단조로 증가하며(Figure 2(a)), 오답의 72%가 margin < 0.5에 몰려 있다. 즉 문제는 오판이 아니라 저신뢰 구간의 크기이므로, 아는 것은 건드리지 않고 모르는 것만 다시 판단하게 하면 된다.

보정 모듈은 Qwen, adapter, structured head를 모두 동결하고 head의 내부 증거만 입력으로 받는다. 76차원 feature는 표준화된 direct/unary/pair 후보 점수 $24\times 3 = 72$개에 margin, 정규화 엔트로피, identity 확률, short flag(공백 분리 24단어 이하)의 스칼라 4개를 이어 붙인 것($72 + 4 = 76$)이며, LayerNorm–Linear(64)–GELU–Dropout–Linear(24)를 거친 잔차 r이

$$z_\pi^{\mathrm{corr}} = z_\pi + g_i \cdot \sigma(s_r) \cdot r_\pi, \qquad g_i = \sigma\big((0.75 - m_i)/0.20\big)$$

로 더해진다(m_i는 행 i의 margin). 학습 파라미터는 6,641개다. 마지막 층을 0으로 초기화해 학습 시작 시점에 보정 후 logits가 기반과 정확히 같고, 학습된 공유 scale은 $\sigma(s_r)$ 기준 0.250에서 0.321로 이동했다. 실제 크기로 보면 margin 0.20인 행에서는 게이트×scale

Table 3: 검증 오답 분해(보정 전)와 보정 모듈 설계의 대응. W4에는 의도적으로 대응 설계를 두지 않아 5.2절 검증의 대조군이 된다.

진단된 약점	약점으로 판단한 근거 (오답 460건 분해)	대응 설계
W1 저신뢰 구간이 넓음	margin<0.5가 332건(72%)	margin·entropy 특징, 게이트 임계 0.75
W2 identity 과예측	false identity 176건(38.3%)	identity 확률 특징
W3 짧은 캡션	≤24단어가 318건(69%)	short flag 특징, 손실 가중 1.25
W4 3↔4 경계 혼동	인접 swap 154건 중 59건	없음 (대조군)

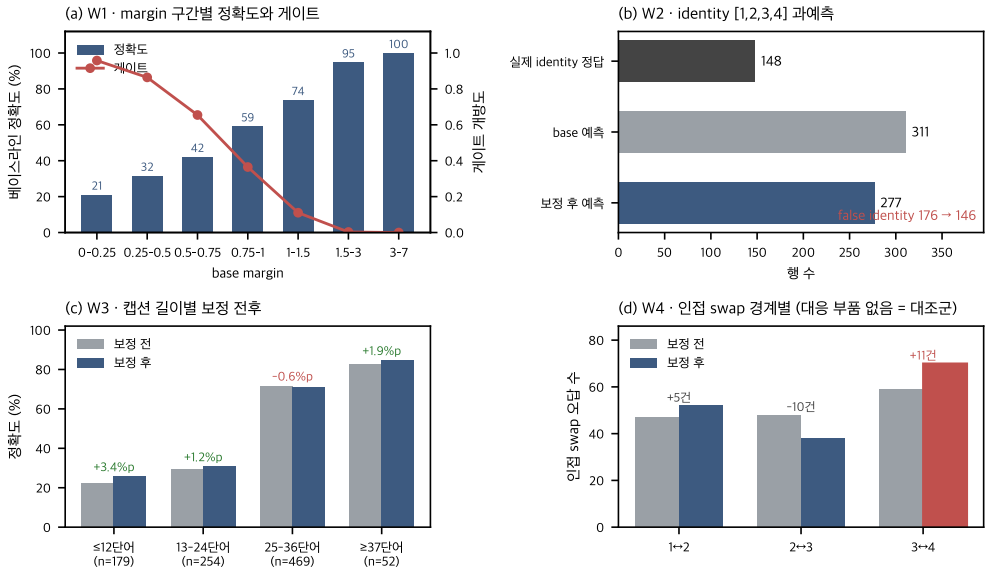

Figure 2: Table 3에서 진단한 네 약점 축의 분포와 보정 전후(고정 검증 954건 실데이터). (a) 정확도가 margin에 완전 단조이므로 저신뢰 구간만 여는 게이트가 정당화되며, 임계 0.75는 42→59% 전이 구간에 놓인다. (b) prior가 만든 identity 과예측(311건)을 보정이 277건으로 되돌린다. (c) 목표 구간(≤24단어)이 개선되고 인접 구간이 소폭 하락하는 의도된 트레이드오프. (d) 대응 설계가 없는 W4만 악화된다. identity에서 풀려난 예측의 일부가 마지막 두 프레임의 혼동으로 이동한 것으로, 개선이 우연한 전반 향상이 아니라 설계된 방향의 이동임을 보인다.

이 약 0.30이라 ±0.6 규모의 잔차가 1위와 2위를 뒤집을 수 있는 반면, margin 1.50에서는 같은 잔차의 영향이 약 ±0.004로 억제된다. 손실은 short 행 1.25배 가중 CE에 margin ≥ 1.0 행의 고신뢰 KL 보존항(가중 2.0)과 잔차 L_2 항(10^{-4})을 더한다. 게이트 임계 0.75는 Figure 2(a)의 정확도 곡선이 42%에서 59%로 꺾이는 전이 구간에서 취했고, 개발 캐시의 그룹 단위 내부 audit 1,124건에서 0.50/0.75/1.00을 비교해(각 528/532/530건 정답) 0.75와 14 epochs를 확정하였다. 이 선택은 리더보드가 아니라 내부 audit로만 이루어졌다.

4 학습 및 추론 과정

전체 lifecycle은 Table 4와 같이 다섯 단계로 고정되어 있다. 개발 단계에서는 train 5,607건으로 학습하고 검증 954건에서 epoch·TTA·prior를 선택했다. 레시피 확정 후 pinned base(NF4 4-bit, BF16 연산)에서 제공 train 9,535건 전체로 새로 학습했다. 구성은 language LoRA r32/α64와 vision 마지막 2블록 LoRA, batch 1×누적 16, paged AdamW 8-bit, cosine 스케줄(warmup 5%), seed 43이다. 학습률은 LoRA 1×10^{-4}, head 5×10^{-4}, weight decay 0.01이고, 이미지 해상도 예산은 행당 최소 100,352에서 최대 200,704픽셀로 고정했다. 증강은 결정론적 균형 입력 셔플로, 24와 서로소인 계수로 샘플 순번과 epoch에 따라 입력 순서를 배정해 연속 샘플이 24개 순열을 고르게 순회하게 하고 라벨을 그 순서에 맞게 재매핑한다. 실제 구현은 샘플마다 서로 다른 두 순서를 생성해 각 0.5 가중으로 손실을 계산하므로, 한 샘플이 한 갱신 단계에서 두 가지 시점 배열을 동시에 학습한다. 제시 순서라는 무의미한 신호를 학습하지 못하게 하는 장치다. 개발 run의 best epoch은 1이었으나 최종 전체학습은 팀이 별도로 고정한 2 epochs를 사용하며, 두 값은 설정 파일에 분리 기록되어 있다. 이후 동결된 모델을 추론 모드로 한 번 실행해 전체 train의 component logits를 캐시하고, 27B 백본 재실행 없이 보정 모듈만 14 epochs 학습해(batch 256, 학습률 8×10^{-4}) 기존 head와 함께 하나의 augmented head로 패키징했다. 추론은 위치 균형 TTA 4뷰(각 프레임이 네 위치에 정확히 한 번씩 등장)를 canonical 순서로 복원해 평균하고, train에서만 계산한 class prior(강도 1.0)를 더한 뒤 argmax한다. 예선 test 3,884건 전체를 A100-40GB 1장에서 3시간 44분에 처리했고 peak 예약 메모리는 17.24 GiB였다.

Table 4: 개발에서 최종 모델까지의 학습 lifecycle. 개발 split은 레시피 선택에만 사용하며, 최종 base와 보정은 각각 9,535건 전체에서 새로 학습한다.

단계	데이터 / view	학습 대상	고정된 결정
Base 개발	5,607 train / 954 val	LoRA + structured head	dev best epoch 1
Base 최종	9,535 / 균형 셔플	새 base의 LoRA + head	2 epochs
보정 개발	5,607 TTA2 / 954 TTA4 cache	잔차 6,641 params	임계 0.75 · 14 epochs
보정 최종	9,535 TTA1 cache	잔차만	prior 1.0 · 14 epochs
최종 추론	입력마다 TTA4	—	canonical 평균 → prior → argmax

5 실험 결과와 분석

5.1 주요 결과와 ablation

Table 5는 예선 test 819건(public 573 + private 246)에 대한 주요 구성의 실측이다. 백본 세대 교체가 가장 큰 단일 기여였고, 잔차 보정은 public +4건, private +1건을 더해 합산 761건으로 최고를 기록했다. epoch 3까지 완전 수렴시킨 변형(751건)과 hard-negative 항을 추가해 재학습한 변형(750건)이 모두 epoch 2 체크포인트(756건)를 넘지 못해, 완전 수렴이 유리하다는 증거는 관측되지 않았다.

고정 검증에서 보정의 통제 비교는 두 기준선으로 보고하며, 두 비교 모두 개발 캐시에서 prior 강도 0.75 조건으로 수행되었다. 저장 경로·정밀도 차이를 제거한 same-cache 기준선 대비 489→501(수정 38건, 훼손 26건, net +12)로 사전 등록 게이트를 통과했고, 보관된 공식 기준선 대비로는 494→501(net +7)로 더 엄격한 +8 임계값에 한 건 미달했다. 두 범위 모두에서 Macro EM(24개 정답

Table 5: 예선 test 819건 ablation(정답 행 수). 최종 모델의 리더보드 점수는 public 0.93542, private 0.91463이다.

구성	public(573)	private(246)	합산(819)
Qwen3-VL-32B full fit	522	214	736
Qwen3.5-27B, epoch 3 (완전 수렴)	526	225	751
Qwen3.5-27B + hard-neg 재학습	526	224	750
Qwen3.5-27B, epoch 2 (기반 모델)	532	224	756
+ 저신뢰 잔차 보정 (최종)	**536**	**225**	**761**

class별 EM의 단순 평균)이 상승하고 최빈 예측 비중이 감소했으므로, 자동 통과가 아니라 명시 기록된 수동 승격 결정을 적용했음을 밝힌다. TTA는 1/2/4뷰에서 검증 EM 47.69/49.69/51.78%로 4뷰가 +4.09%p를 더했고, 12뷰 확장은 별도 표본에서 이득이 없어(72.0 대 72.5%) 4뷰로 고정했다.

5.2 보정이 의도한 구간에서만 작동했는가

보정이 설계 의도대로 작동했는지를 보정 전후의 행 단위 대조로 확인했다. 예측이 바뀐 169건 중 168건(99.4%)이 게이트 개방 구간(게이트 ≥ 0.8, 431건)에서 발생했고, 차단 구간(< 0.3, 335건)에서는 한 건도 바뀌지 않았다(Table 6). 기반 모델이 90.9%의 정확도를 보이던 margin ≥ 1.0의 318건은 예측 변경과 훼손이 모두 0건으로 완전히 보존되었다. zero 초기화, margin 게이트, 고신뢰 KL 보존항의 결합이 실측으로 확인된 것이다.

Table 6: 게이트 구간별 예측 변경. 변경 169건 중 168건(99.4%)이 개방 구간에서 발생했고 차단 구간은 0건이다. 약점 축별 보정 전후는 Figure 2에 있다.

게이트 구간	행수	예측 변경
개방 ($\geq$0.8)	431	168
부분 (0.3–0.8)	188	1
차단 ($<$0.3)	335	**0**

다만 총 이득의 성격은 정직하게 기술할 필요가 있다. 오답→정답 39건과 정답→오답 32건의 기반 margin 중앙값은 각각 0.156과 0.110으로, 양쪽 모두 사실상 동률 구간이다. 즉 보정의 실체는 확실한 오류의 교정이 아니라 거의 동률인 구간에서의 미세한 승률 우위이며, 그래서 EM 이득(+7건)보다 Macro EM(+1.19%p)과 최빈 예측 비중(−3.56%p) 같은 분포 지표의 개선이 훨씬 크다. 또한 해소된 false identity 34건 중 정답에 도달한 것은 5건뿐이고 29건은 다른 오답으로 이동했다. identity 편향은 걷어냈지만 그것이 곧 정답 발견으로 이어지지는 않았다.

5.3 오류 분석

오류 유형은 인접 2-swap 33.5%, 3-순환 32.4%, 전면 뒤섞임 19.8%, 비인접 2-교환 11.7%, 완전 역순 2.6% 순이며, 인접 swap이 혼동한 시점 쌍은 3↔4가 38%로 가장 많고 보정 후 오히려 44%로 늘었다(경계별 전후는 Figure 2(d)). 한편 margin 1.5 이상에서 틀린 행은 954건 중 6건뿐이며 전부 인접 2-swap이었다. 확신을 갖고 크게 틀리는 경우는 사실상 없다는 뜻으로, Figure 2(a)의 캘리브레이션과 정합한다. 캡션 길이는 지배적 요인으로, 12단어 이하 구간 정확도는 22.3%에 그친 반면 37단어 이상은 82.7%였다(Figure 2(c)). 오답이 서로 다른 두 한계 유형에서 나온다는 점도 확인했다. 프레임 간 시각 유사도(96×96 픽셀 평균절대차)로 검정하면 긴 캡션 오답은 시각적으로 유사한 프레임과 유의하게 연관되지만(Cohen $d = 0.50$, $p < 10^{-4}$), 짧은 캡션 구간에서는 그 연관이 사라진다($d = 0.04$, $p = 0.73$). 즉 전자는 지각 해상도의 문제(지각 한계), 후자는 캡션 정보 부재의 문제(정보 한계)로 서로 다른 처방을 요구한다. 짧은 캡션의 정보 부재는 image-only·text-only 진단(각각 3.5%, 우연 수준 4.2%)과 시간 표현이 없는 캡션의 identity 비율이 전역 prior와 동일한 15.5%라는 관찰로도 뒷받침된다. 한편 class prior(identity에 +1.31, 강도 1.0)는 평균 성능을 올리지만 저신뢰 행을 identity로 흡수하는 부작용을 만들며(Figure 2(b)), 전역 강도 조정으로는 해결되지 않아 행별 조건부 보정의 필요성이 도출되었다.

5.4 사례 연구

검증의 실제 행으로 보정의 성공·실패와 두 한계 유형의 실물을 보인다. 프레임은 모두 대회 제공 train 데이터이며 모델에 입력된 순서 그대로다.

Figure 3은 5.3절의 두 한계 유형을 나란히 놓은 것이다. 위 **NWKfDF**(31단어, margin 0.008)는 네 설경 프레임이 육안으로도 거의 구분되지 않는 지각 한계 사례로, 기반이 이미 정답 $[2, 1, 4, 3]$이었으나 보정이 identity로 바꿔 훼손했다(W2 처방의 오발). 낮은 margin이 곧 오답을 뜻하지 않으며 고정 게이트와 identity 보정이 모든 사례에서 안전하지는 않다는 한계다. 아래 **zvILfH**(7단어, margin 0.023)는 반대로 프레임들이 뚜렷이 다르지만 캡션 "The man lassos and ties it up."이 순서에 대해 아무 정보도 주지 않아, 보정 전($[2, 1, 3, 4]$) 후($[3, 1, 4, 2]$) 모두 정답 $[3, 1, 2, 4]$에 이르지 못한 정보 한계 사례로, 짧은 캡션 약점 W3의 전형이다.

Figure 4 위의 **ssSIgO**(13단어, margin 0.076)는 보정의 성공 구간이다. 기반은 identity $[1, 2, 3, 4]$를 답했으나 실제 정답은 완전 역순 $[4, 3, 2, 1]$(노인이 콘 코스를 제시 순서의 반대로 통과하는 장면)이며, 보정이 identity 확률·margin 신호로 이를 복구했다. W2 진단이 겨냥한 정확히 그 지점이다. 마지막으로 같은 그림 아래의 **LVOSrj**(40단어, margin 2.31)는 남은 약점 W4의 실물이다. 전체 흐름은 확신을 갖고 맞혔으나 시간 3·4위에 해당하는 두 프레임의 선후만 뒤집었고($[1, 2, 4, 3] \rightarrow [1, 2, 3, 4]$), margin이 높아 게이트가 닫히므로 보정이 손대지 못한다. 게이트 값으로 보면 ssSIgO는 0.97로 열려 있었고 LVOSrj는 0.0004로 사실상 닫혀 있어, 같은 게이트 식 하나가 두 사례의 상반된 결과를 결정했다.

5.5 시도했으나 채택하지 않은 기법

모든 판정은 리더보드 제출 없이 고정 검증에서 이루어졌다. 생성형 verbalizer SFT는 분류와 등가(49.79 대 49.58%), 단일 run 내 SWA(학습 후반 checkpoint들의 가중치 평균)는 완전 동점($p = 1.0$), 12뷰 TTA는 4뷰 대비 무효, 고해상도 추론은 −0.52%p, "margin이 낮으면 identity" 규칙은 전 임계값에서 손해(저신뢰 행의 실제 identity 비율 10%는 prior 15.5%보다 오히려 낮음), 캡션 시간 표현 규칙은 정보량 0으로 각각 기각했다. hard-negative 항을 추가한 재학습은 예선 실측에서 기반 모델보다 낮아(Table 5) 채택하지 않았다.

Figure 3: 두 한계 유형의 실물 대비(대회 제공 train 프레임, 입력 순서). 위 NWKfDF는 네 프레임이 거의 구분되지 않는 지각 한계(긴 캡션 오답이 시각 유사도와 강하게 연관되는 유형, $d = 0.50$)이며, 저신뢰(W1) 구간에서 identity 보정이 정답을 훼손한 W2 처방의 오발 사례다. 아래 zvILfH는 프레임이 뚜렷이 다른데도 7단어 캡션이 순서 정보를 주지 않는 정보 한계(시각 유사도와 무관한 유형, $d = 0.04$)이며, 짧은 캡션 약점 W3의 전형이다.

Figure 4: 보정의 경계. 위(**ssSIgO**)는 보정 성공 사례로, 정답은 완전 역순 $[4, 3, 2, 1]$인데 기반이 identity를 답했고(margin 0.076) 보정이 복구한 W2의 표적 지점이다. 아래(**LVOSrj**)는 남은 약점 W4로, margin 2.31의 고신뢰에서 시간 3·4위 두 프레임(입력 3·4)의 선후만 뒤집혔으며 게이트가 닫혀 보정 대상이 아니다.

6 방법론의 장점과 한계, 향후 방향

장점. 출력이 구조적으로 항상 유효한 순열이라 후처리가 불필요하고, CE가 평가 지표를 직접 최적화한다. 보정은 별도 모델의 예측 결합이 아니라 동일 head의 내부 증거를 다시 읽는 잔차 항이므로 단일 모델·단일 checkpoint 규정 안에 있으며, 캐시 기반이라 27B 재실행 없이 반복 실험이 가능하다. 진단·설계·검증 사슬 덕분에 각 설계 요소의 존재 이유가 실측 수치로 소급된다.

한계. (1) 공식 기준선 대비 net +7은 사전 보수 기준 +8에 미달했고(양측 부호검정 $p \approx 0.48$), 이득의 대부분이 동률 구간의 승률에서 나온다. (2) 검증셋의 짧은 캡션 비율(45.4%)이 실제 평가 분포와 달라 개선 폭 추정에 왜곡이 있으며, 분할이 캡션 길이를 층화하지 않은 것이 원인이다. (3) short flag는 단어 수 하나의 지표로 의미적 모호성과 시각 난이도를 구분하지 못한다. (4) 3↔4 경계는 시각 유사도로 설명되지 않는 잔존 약점으로 보정 후 오히려 악화되었다. (5) 보정 게이트가 기반 모델이 학습에 사용한 데이터의 캐시로 보정되어 margin 분포가 낙관적일 수 있다. (6) pHash 그룹화는 휴리스틱이며 하나의 고정 검증만 사용해 개선 폭의 분산을 측정하지 못했다. (7) 최종 전체학습 모델은 검증 954건을 학습에 포함하므로 독립 검증 수치가 존재하지 않으며, 보정의 개별 특징·손실 항의 독립 기여도 분리 측정하지 못했다.

향후 방향. 3↔4 경계에는 프레임 토큰 풀링과 경계 인지(boundary-aware) 인접 관계 branch를, 저정보 캡션에는 캡션 모호성 신호의 세분화와 함께, 방향 어휘를 보존하는 규칙 기반 캡션 축약으로 학습 난이도를 높이는 curriculum 및 짧은 캡션 oversampling을 검토한다. 규칙 기반으로 한정하는 이유는 생성 모델을 이용한 데이터 변형이 규정상 허용되지 않기 때문이다. 각 항목은 remove-one feature ablation과 grouped bootstrap 반복 분할을 포함해, 동일한 그룹 분할 검증에서 독립적으로 확인하는 것을 판정 기준으로 삼는다.

7 재현성과 규정 준수

대회 제공 train만 학습에 사용했고 외부 데이터·외부 추론 API·생성형 증강·pseudo-labeling·모델 앙상블을 사용하지 않았다. 최종 추론은 pinned base(공개 시점이 가중치 컷오프 2026-05-31 이전인 revision) 하나, LoRA adapter 하나, augmented head 하나를 로드하며, 동일 checkpoint의 입력 순서 TTA 4뷰만 사용한다. 같은 단일 checkpoint를 도메인이 다른 외부 검증 test에도 변경 없이 그대로 적용하였다(공개 구간 50건 기준 0.78). 고정 분할·최종 레시피·승격 판단 기록(미달 사실 포함)은 기계 판독 가능한 설정 파일로 저장소에 고정되어 있고, 재학습·보정·추론 스크립트는 이 기록이 없거나 모순되면 실행을 거부한다. 최종 제출 CSV의 SHA-256(`0fee7d90…`)과 adapter·augmented head의 SHA-256을 저장소 README에 기재해, 평가자가 내려받은 산출물이 본 보고서의 실험과 동일한 바이트임을 대조할 수 있게 하였다.

참고문헌

[1] Qwen Team, *Qwen3.5-27B*, Hugging Face repository, `huggingface.co/Qwen/Qwen3.5-27B`.
[2] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," 2021.
[3] T. Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," 2023.
[4] H.-Y. Lee et al., "Unsupervised Representation Learning by Sorting Sequences," 2017.
[5] I. Misra et al., "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification," 2016.