텍스트로 풀어보는 장면의 재구성 보고서_HHH

1. 전체적인 방법론 및 모델 구조

1.1 문제 정의

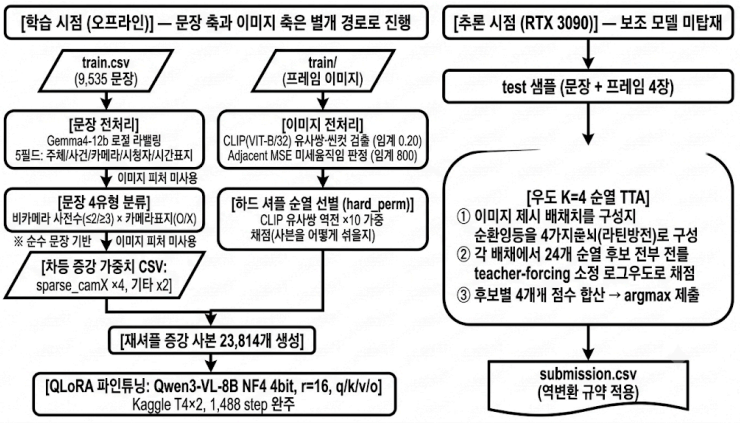
뒤섞인 4개의 비디오 프레임과 텍스트 캡션이 주어졌을 때, 각 프레임의 실제 시간적 순서 `[n,_n,_n,_n]`을 복원하는 문제, 어려운 이유는 아래 4가지임.

1. **평가 지표의 엄격성**: 정답 여부는 완벽히 맞혔는지로만 판단하며 부분 점수는 없음. 24가지 순서 중 하나만 정답이라 무작위로 찍었을 때 정답률은 1/24 ≈ 4.2%.
2. **Zero-shot의 완전한 무력함**: 후보 VLM 5종(Qwen2-VL-2B, Qwen2.5-VL-3B/7B, Qwen3-VL-2B/4B)의 Zero-shot 추론 결과, 순서가 뒤섞인 프레임에 대한 정확도가 0.8~4.4%로, 이는 사실상 무작위 수준
3. **VLM의 구조적 위치 편향**: 멀티 이미지 VLM은 "먼저 제시된 프레임 = 시간상 먼저"라는 강한 편향을 갖고 있어, 제시 순서를 그대로 베낀 identity(`[1,2,3,4]`) 오답이 반복됨.
4. **문장의 시간적 모호성 편차**: 시간 표지가 없는 문장은 정확도가 40.5%p 떨어지고, 카메라 표현 유무만으로도 26.9%p가 갈리는 등, 문장이 시간적으로 얼마나 명확한가에 따라 난이도가 극단적으로 달라짐.

1.2 실험 공통 규약 및 용어 정의

용어	정의
holdout_300	train.csv 9,535건에서 분리한 고정 평가셋 300건(셔플 샘플 252건 + 비셔플 샘플 48건)
acc_shuffled	holdout 중 셔플 252건의 정확도, 비셔플(1,2,3,4) 포함 총점은 판정에 사용하지 않음.
acc_identity	비셔플 48건의 정확도, 과억제 감시용 보조 지표
재셔플 증강 ×N	동일 샘플의 프레임 제시 순서를 재배열하고 정답 라벨을 자동 재계산한 사본을 N개 학습에 주입
미니 스크리닝	1,000샘플·300스텝 축소 학습(≈1.5시간)으로 후보 기법을 1차 선별하는 절차. 통과한 후보만 풀 학습으로 승급

1.3 End-to-End 파이프라인 개요



2. 제안한 아이디어의 동기와 핵심 기여

2.1 베이스라인의 한계 분석

제공된 베이스라인은 세 가지의 한계를 지닌다.

1. 베이스라인 코드 기반으로 추론한 VLM 5종의 Zero-shot 정답률이 0.8%~4.4%에 불과하여 무작위 추첨(1/24 ≈ 4.2%) 수준에 그침.
2. 정답이 24개 순열 후보로 한정된 문제임에도 주관식 토큰 생성을 채택. 이로 인해 첫 토큰 결정에 전체 결과가 좌우되는 한계와 포맷 오류에 따른 파싱 오답이 빈번.
3. 시각 내용과 무관하게 "첫 이미지 = 1번 프레임"으로 오인하여 제시 순서 그대로 출력하는 경우가 빈번함.

2.2 핵심 아이디어

1. 언어학 이론 기반 문장 4유형화 및 차등 증강

- Vendler 동사 상 이론, TimeML, SDRT 담화 구조 이론을 바탕으로 로컬 Gemma4-12b로 9,535문장을 5필드(주체·사건·카메라·시청자·시간표지)로 전량 라벨링.
- 문장 내 사건 수, 카메라 표지 조합의 4개 문장 유형을 확정하고, 최약점 유형(`sparse_camX`)에 4배의 증강 배수를 집중 부여함.

2. 우도 기반 추론

- 24개 순열 후보 전체의 로그우도를 직접 채점하고, 라틴방진 4회전 배치의 채점 결과를 합산하여 정답 순열을 도출함.
- 별도 학습 없이 추론 단계 개선만으로 **Holdout +4.76%p, Private +4.1%p**의 프로젝트 단일 최대 성능 향상을 달성함.
- 자원 제약 환경에서의 단계적 스케일링

  2B(베이스라인 레시피 확립) ➔ 4B(프롬프트·증강 최적화 검증, +7.3%p) ➔ 8B(최종 완주)로 모델 크기를 순차적으로 확장, 각 체급에 최적화된 학습 세팅을 도출.

3. 학습 및 추론 과정, 주요 기법 (Training & Inference / Proposed Methods)

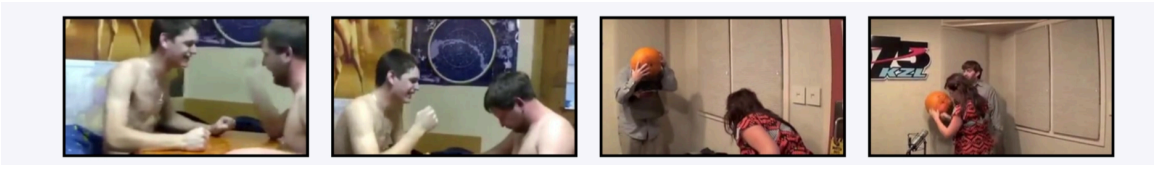
3.1 데이터 전처리 (1): 문장 성분 라벨링

정규식 기반 1:1 매칭은 상적 변화처럼 문맥에 따른 변화를 처리하지 못함. 이를 해결하기 위해 **로컬 Gemma4-12b(WSL Ollama)**로 전체 9,535건을 전량 라벨링.

- **추출 스키마**: JSON 5필드 (`subjects`, `events`, `camera_language`, `viewer_language`, `temporal_markers`)
- **모델 선정**: Gemma4-e4b는 사건 과소 분할로 인한 누락, 12b 모델 채택.
- **품질 검증**: 파싱 성공률 100%, 카메라 라벨 정규식 대조 92.1%

3.2 데이터 전처리 (2): 이미지 분석

- **CLIP(ViT-B/32) 씬컷 검출:**
  - 이미지의 의미적 유사도 측정을 위해 clip을 채택, 이미지 유사도 기준값을 0.2로 잡았을 때 오탐지율이 4.8% 미만으로 가장 낮음.
  - 4개 프레임에서 나오는 6개(${}_4C_2$) 쌍 중 5개 이상이 비슷하면 장면 전환 0회로 보는 느슨한 기준을 적용했고, 그 결과 전환 0회 27.7% / 1회 39.5% / 2회 21.4% / 3회 11.4%로 나타남.



- **탐색 공간 축소**
  - 비슷한 장면끼리 분석한 결과, 1회 전환(`[{0,1},{2,3}]`)시 감지 시 유효 순열이 24 → 4개로 **83.3% 축소**
  - 학습 단계의 하드 셔플 순열 선별(§3.4) 및 힌트 프롬프트 실험(§3.5)에 활용

3.3 문장 4유형 정의와 holdout 실측 확정

- 문장 내 사건 수, 카메라 표지 조합의 4개 문장 유형을 확정함.

	문장 내 사건 수 ≥ 3	문장 내 사건 수 ≤ 2
카메라 표지 O	dense_camO	sparse_camO
카메라 표지 X	dense_camX	`sparse_camX`

- 문장 유형별 Holdout 결과값을 바탕으로 보완이 필요한 데이터 유형을 결정

문장 유형	Holdout 정확도	Train 분포	전처리 전략
sparse_camX	**21.6% (n=74)**	**2,755 (29%)**	**차등 증강(×4)**
sparse_camO	48.7% (n=39)	1,114 (12%)	×2
dense_camX	52.7% (n=74)	3,084 (32%)	×2
dense_camO	73.8% (n=65)	2,582 (27%)	×2

3.4 재셔플 증강

재셔플 증강은 동일한 샘플에 대해 프레임 이미지의 제시 순서를 무작위로 재배열하고, 변경된 위치에 맞게 정답 라벨을 자동 재계산한 사본을 데이터셋에 주입하는 기법임.

동일 데이터의 제시 순서를 재배열한 사본을 2배 주입했을 때 1배 대비 Holdout 정답률이 +6.35%p 상승하였으며(exp07 48.41% vs exp06 42.06%), 이를 §3.3의 최약점 유형 차등 배수(sparse_camX ×4)와 결합함.

4B Greedy 통제 실험(exp20)에서 Holdout +0.4%p 대비 Public -1.1%p의 역효과가 관측되어 1차 기각되었으나, 최종 8B 모델은 마감 일정상 해당 레시피의 977 Step 체크포인트를 완주시키고 우도 K=4 TTA 추론을 결합하여 셔플 방식의 불리함을 상쇄하고 최종 Private 0.86991를 달성함.

3.5 프롬프트 엔지니어링

프롬프트 구조 실험 15종 수행 결과, 역할 부여나 스토리텔링 프레이밍 등은 단순 노이즈로 작용하였으며 문장(캡션)의 위치 이동(v5_reorder)만이 유의미함.

- v5_reorder는 같은 문장을 지시문 뒤로 옮김. 문장이 답 바로 직전에 있으면 모델이 이미지 4장을 모두 인코딩한 후에도 문장 맥락을 손실 없이 참조하여 매핑 수행.

```
Look at the 4 images above labeled Image 1 to Image 4. Determine the correct
chronological order of these images to match the sentence below.
Sentence: "{sentence}"            ← 문장을 답 출력 직전(맨 뒤)으로 배치
Provide the answer ONLY as a Python list of integers. Example: [1, 2, 3, 4]
```

- CLIP으로 감지한 장면 전환 수치는 다음과 같이 프롬프트에 힌트로 추가됨.

```
[Hint] Visual analysis indicates these 4 images contain {cuts} distinct scene cuts. Images {i} and {j} are near-duplicate frames.
```

- CoT 지도학습 제외: 단계별 추론 과정을 출력하게 하는 CoT 지도학습은 4회 연속 실패로 영구 기각함.

3.6 파인튜닝 파라미터 설정

항목	설정	비고
LoRA	r=16, α=32, targets=[q,k,v,o]_proj	비전타워 동결, LLM Attention 4모듈
양자화	2B: bf16 / 4B·8B: NF4 4bit QLoRA	8GB 로컬 GPU에서 4B 학습 가능
최적화	AdamW + Cosine + Warmup 3%, lr 1e-4	
배치 구성	grad_accum 16, epochs 1 — 23,814 증강 항목 → 1,488 opt steps	100스텝마다 중간 저장
입력 해상도	max_pixels: 로컬 307,200(640×480) / 8B 512×384	Kaggle T4 16GB 여유로 8B에서 상향
학습 인프라	로컬 RTX 5060 8GB (2B/4B) / Kaggle T4×2 (8B)	

3.7 추론 기법

1. 우도 스코어링

모델 추론의 정답은 4! = 24개의 객관식 순열 공간 내에서 결정됨. 문장과 프레임 4장으로 구성된 프롬프트 x가 주어졌을 때, 정답 후보 순열 a ∈ S_4에 대응하는 답변 문자열 y(a) = `"[a1, a2, a3, a4]"`를 토큰열 (t1, ..., tL)로 분해, 이후 토큰을 입력에 넣어 건부 로그 우도(Log-Likelihood) s(a)를 계산함:

$$s(a) \;=\; \sum_{k=1}^{L} \log P_\theta\big(t_k \,\big|\, x,\; t_1, \ldots, t_{k-1}\big)$$

공통 입력 프롬프트 x의 Key-Value(KV) 캐시를 1회 사전 계산한 뒤 24개 후보가 공유하므로, 후보당 추가 연산 비용은 짧은 답변 토큰열 1회의 Forward 패스에 불과함.

2. K=4 순환 배치 TTA (라틴방진 합산)

이미지 제시 배치 π를 순환이동 4가지

$$\Pi = \{(1,2,3,4),\; (2,3,4,1),\; (3,4,1,2),\; (4,1,2,3)\}$$

로 바꿔가며 (a)의 우도 채점을 반복. 이 4개 배치는 모든 이미지가 4개 제시 슬롯을 정확히 1회씩 방문하는 최소 균형 집합(라틴방진)임. 배치 π 하에서 원본 좌표의 후보 a를 채점하는 문자열은 $y_\pi(a) \;=\; \big[a_{\pi(1)},\, a_{\pi(2)},\, a_{\pi(3)},\, a_{\pi(4)}\big]$로 재매핑 후 최종 점수와 제출:

$$\mathrm{Score}(a) \;=\; \sum_{\pi \in \Pi} s_\pi(a), \qquad \hat{a} \;=\; \arg\max_{a \in S_4} \mathrm{Score}(a)$$

우도 스코어를 쓰는 이유: 위치 편향으로 인한 노이즈 점수는 4회 회전 과정에서 서로 다른 후보로 분산되어 상쇄되는 반면, 실제 시각적 근거 점수는 동일한 정답 후보 a에 4회 연속 누적되므로 편향을 구조적으로 완벽히 제거함.

3. 실측 효과

- **Holdout 검증셋 개선**: 로컬 4B 모델(exp17) 적용 시 acc_shuffled가 **51.98% ➔ 57.14% (+5.16%p)**로 대폭 상승하여, 학습 없이 추론 레버 단독으로 강력한 성능 향상을 입증함.
- **Private 리더보드 도약**:

모델 및 학습 상태	생성 방식	Private 점수	비고 (성능 변동)
8B 977 Step	Greedy	0.81300	-
8B 977 Step	TTA 적용	0.85365	+4.065%p (Greedy 대비 향상)
8B 1,488 Step (완주)	TTA 결합	**0.86991**	+5.691%p (Greedy 대비 향상)

---

4. 실험 결과 및 분석 (Experiments & Analysis)

4.1 실험 프로토콜

모든 통제 실험은 §1.2의 공통 통제 규약(고정 평가셋 holdout_300 활용, 셔플 정확도 주 지표 채택, ±4%p 오차범위 노이즈 밴드 설정)을 준수하여 수행함.

4.2 종합 결과 마스터 테이블

모델	프롬프트	증강 및 셔플 레시피	추론 방식	acc_score	Public 점수	Private 점수	비고 및 비평
2B	v1_list	미적용 (Zero-shot)	Greedy	3.17%	0.76265	0.76829	exp01 (무작위 수준 4.2%)
2B	v1_list	재셔플 ×2 / 무작위	Greedy	48.41%	0.76614	0.73983	exp07 (재셔플 증강 채택)
2B	v5_reorder	재셔플 ×2 / 무작위	Greedy	48.02%	0.77486	0.77642	exp14 (문장 후치 양성 반응)
2B	v1_list	sparse_camX ×4 / 무작위	Greedy	48.02%	0.78359	0.76016	exp16 (약점 유형 차등 증강)
4B	v5_reorder	sparse_camX ×4 / 무작위	Greedy	51.98%	0.85689	0.82520	exp17 (체급+프롬프트 대약진)
4B	**v5_reorder + 힌트**	**sparse_camX ×4 / 무작위**	**Greedy**	**51.50%**	**0.85514**	**0.85365**	**v10 (4B Private 최고)**
4B	v5_reorder	sparse_camX ×4 / 하드 셔플	Greedy	52.38%	0.84642	0.83739	exp20 (Public -1.1%p 기각)
8B (977)	v5_reorder	sparse_camX ×4 / 하드 셔플	Greedy	—	0.83944	0.81300	8b 977 Step 체크포인트
8B (977)	**v5_reorder**	**sparse_camX ×4 / 하드 셔플**	**우도 K=4 TTA**	**—**	**0.88830**	**0.85365**	8b **TTA 단독 적용 (+4.1%p)**
8B (1488)	**v5_reorder**	**sparse_camX ×4 / 하드 셔플**	**우도 K=4 TTA**	**54.67%**	**0.88307**	**0.86991**	**최종 제출본**

4.3 Ablation Study(각 기법의 기여도)

레버	대조 실험	독립 기여	판정
파인튜닝 자체	zero-shot 0.8% → exp01 45.63%	+44.8%p	학습 신호 존재
재셔플 증강 ×2 (vs ×1)	exp06 42.06% → exp07 48.41%	+6.4%p (holdout)	채택
v5_reorder 프롬프트	exp07 0.766 → exp14 0.775	+0.9%p (Public)	채택
sparse_camX 타깃 증강	exp07 0.766 → exp16 0.784	+1.8%p (Public)	채택
체급 스케일업 (미니 스크리닝)	미니 2B 12.3% → 미니 4B 25.79%	+13.5%p	4B/8B 승급
4B + v5_reorder (풀 학습, 2변수)	exp16 0.784 → exp17 0.857	+7.3%p (Public)	채택 (최대 학습 레버)
CLIP scene_cuts 힌트	exp17 Private 0.825 → v10 0.854	+2.85%p (Private)	4B Private 최고
하드 셔플	exp17 0.857 → exp20 0.846	−1.1%p (Public)	기각
CoT SFT	exp06 42.06% → exp12 37.7%	−4.4%p 포함 4연패	영구 기각
우도 K=4 TTA	8B 977 Greedy 0.813 → TTA 0.854	**+4.1%p (Private)**	최대 추론 레버
1488 완주	8B 977 TTA 0.854 → 1488 TTA 0.870	+1.63%p (Private)	채택

4.4 977 vs 1488 완주 학습의 데이터 커버리지 인과 분석

8B 모델 학습은 Kaggle 12시간 세션 제한으로 인해 복수 세션으로 나누어 진행되었으며, 977 Step(전체 1,488 Step의 65.7%) 중간 체크포인트를 1차 제출함. 이후 1,488 Step 완주를 성공적으로 마쳤으며, 두 모델 간의 제출본 간 발생한 Private 점수 +1.63%p 상승(0.85365 ➔ 0.86991)의 원인을 데이터 커버리지 관점에서 규명한 사후 진단(Post-hoc) 분석임.

- **학습 커버리지 점검.** 총 증강 사본 23,814개 중 977 step까지 8,182개(34.4%)가 미학습.

유형	배수	총 사본	미학습 사본 (8,182 중)	완전미학습 ID (837 중)	Test 변경률	z-score
dense_camO	×2	5,000	1,754 (21.4%)	**317 (37.9%)**	12.2%	−3.39
dense_camX	×2	5,990	2,067 (25.3%)	**368 (44.0%)**	20.9%	−0.37
sparse_camO	×2	2,136	702 (8.6%)	116 (13.9%)	13.0%	−1.84
sparse_camX	×4	10,688	**3,659 (44.7%)**	36 (4.3%)	52.0%	+7.38

  - ×4 증강 sparse_camX: 완전미학습 확률이 $(0.344)^4 \approx 1.4\%$로 낮아 ID 유실은 36개(4.3%)에 불과했으나, 사본 절대 수량이 많아 완주 구간 미학습 사본의 44.7%(3,659개)를 차지함.
  - ×2 증강 dense 계열: 2개 사본이 모두 유실될 확률이 $(0.344)^2 \approx 11.8\%$로 높아, 완전미학습 ID 837개 중 81.8%가 dense 계열에 집중됨.
- **문장 성분 기준 대조**

성분 축	세부 분류	Train 전체	Train 0점	미학습 비율	Test 전체	Test 변경 수	Test 변경률	관측 현상
사건 개수	ev_4+ (4개 이상)	2,881	365개	43.61% (1위)	469	77	16.42% (최다 변경)	Train 유실 최다 ➔ Test 변경 절대량 최다 관측
사건 개수	ev_3 (3개 사건)	2,614	320개	38.23%	173	38	21.97%	사건 3개 이상 결합 시 미학습 81.8% 형성
사건 개수	ev_2 (2개 사건)	2,433	104개	12.43%	112	35	31.25%	사건 수 감소에 따라 변경 비율 증가
사건 개수	ev_0-1 (0~1개)	1,307	48개	5.73%	65	27	41.54%	단문 핀포인트 씬 매칭 양상 관측
시간표지	mark_0 (표지 없음)	2,423	94개	11.23%	96	49	51.04% (1위)	Test mark_0의 54.2%가 sparse_camX와 중복
시간표지	mark_1 (표지 1개)	2,031	173개	20.67%	108	28	25.93%	중간 수준 변경 비율 관측
시간표지	mark_2+ (2개 이상)	4,781	570개	68.10%	615	100	16.26%	명시적 힌트 존재로 977 시점 이미 안정적

  - 사건이 4개 이상 포함된 고밀도 문장(ev_4+)은 Train 미학습 유실 1위(43.61%, 365개 ID)를 기록함과 동시에, Test 예측 변경 절대 건수에서도 1위(77건)
  - 시간표지가 없는 문장(mark_0)은 Test 변경률이 51.04%(49/96건)로 가장 높게 나타남. 교차 대조 결과, 이는 mark_0 자체의 미학습 때문이 아니라 Test mark_0 96건 중 54.2%(52건)가 ×4 증강이 적용된 sparse_camX 유형에 중복 속해 있어 발생한 현상임
- **분석 결론**
  1. 단일 유형 쏠림을 완화하기 위해 **sparse_camX 집중 배수를 적정 재배분하는 한편, 완전 미학습 유실의 81.8%가 집중된 고난도 dense 계열과 ev_4+ 고밀도 문장의 증강 배수를 ×2 → ×3 sim 4로 집중 상향**하여 조기 중단에 대비한 커버리지 대비 구조를 구축해야 함.
  2. **유형 축을 [dense/sparse] × [camO/X] × [mark0/1+] 8유형으로 세분화하여 최고 난이도 니치인 dense_camX_mark0에 최우선 배수를 부여**하고, 에폭 전 구간에 사본이 균등 배치되도록 Stratified DataLoader를 적용해야 함.

---

5. 자원 및 비용 효율성 (Efficiency & Optimization)

5.1 체급별 스케일링 및 추론 연산 효율성 분석

체급	정밀도	학습 VRAM Peak	추론 속도 (greedy)	acc_shuffled	Public	Private	평가
Qwen3-VL-2B (exp16)	bf16	6.4 GB	0.85초/샘플	48.02%	0.784	0.760	2B 체급 한계
Qwen3-VL-4B (exp17)	4bit	3.8~6.2 GB	1.14초	51.98%	0.857	0.825	Public 대약진
Qwen3-VL-4B (v10)	4bit	3.8~6.2 GB	1.22초	51.50%	0.855	**0.854**	자원 효율 최적점
Qwen3-VL-8B (완주)	4bit	11.2 GB	2.16초	**54.67%** [추후]	**0.883**	**0.870**	

- **EDA 기반 파인튜닝 레시피 최적화의 결정적 우위**
  - 4B v10 모델은 8B 대비 VRAM 3.8GB(약 50% 절감) 환경에서 구동 가능하면서도 최종 모델 성능의 **98%(Private 0.85365 vs 0.86991)**에 도달함.
  - 4B v10 Greedy(0.85365) = 8B 977 Step + K=4 TTA(0.85365)로, 잘 설계된 4B 모델이 최강 추론 기법을 결합한 미완성 8B 모델과 동등한 성능.
- **우도 K=4 TTA 추론의 연산 구조 및 시간 예산 검증**
  - 24개 후보 순열을 채점하는 연산은 프리픽스 Key-Value(KV) 캐시 공유를 통해 추가 연산 오버헤드가 사실상 0에 수렴함.
  - 실제 연산 비용의 본체는 이미지 4회 순환 회전(라틴방진) 인코딩 연산이며, RTX 3090 환경 기준 전체 819건 추론에 약 1~2시간만 소요되어 대회 제한 시간(24시간)의 1/10 이내로 완주 가능함.

5.2 구축 비용

- **외부 상용 API 비용: 총 0원** (한도 3만 원). Gemma4-12b 라벨링(~7.3시간), CLIP 피처 추출, MSE 분석 전부 로컬 수행.
- **학습 연산량**: 2B/4B는 소비자급 노트북(8GB), 8B는 무료 Kaggle T4×2. 유료 클라우드 학습 비용 0원.

---

6. 결론 및 한계점 (Conclusion & Limitations)

6.1 결론

본 연구는 사전학습 VLM의 Zero-shot 성능(0.8%)에서 출발 Public 0.88830, Private 0.86991에 도달하는 과정을 엄격한 통제 실험과 실측 데이터에 기반해 입증.

성능 도약의 3대 핵심 동력

1. 언어학 기반 Data-Centric 차등 증강: Vendler 동사 상 이론과 로컬 Gemma4-12b 구문 라벨링을 통해 도출된 최약점 문장 유형(sparse_camX)에 학습 사본을 4배 집중 주입하여 약점 정답률을 보정함.
2. 파인튜닝 레시피 최적화: v5_reorder 프롬프트(문장 후치) + 차등 배수 + 체급 스케일업을 유기적으로 통합하여, 단순 모델 확대보다 EDA 기반의 설정 최적화가 성능에 더 결정적 요소임을 실증함
3. 우도 기반 K=4 순열 TTA: 24개 객관식 답 공간을 전수 채점하고 라틴방진 회전 배치를 합산하는 방식으로 추론 패러다임을 재설계하여, 추가 학습 비용 0으로 Private +4.1%p 도약을 달성함.

연구 수행의 2대 철학적 원칙

1. 실측 데이터에 기반 검증: 고정 평가셋 300건(holdout_300), 셔플 정확도(acc_shuffled) 단독 판정, 단일 변수 통제, ±4%p 오차범위라는 엄격한 기준을 수립하여 CoT 지도학습, 텍스트 힌트 주입, 하드 셔플 등 그럴듯하지만 실제로는 유해했던 아이디어들을 기각함.
2. 학습 내재화 및 단일 모델 추론: Gemma4, CLIP, OWL-ViT 등 전처리 보조 모델의 인사이트를 데이터 증강 가중치로 전환하여 LoRA 파라미터 내부에 내재화함으로써, 평가 규정(단일 모델 구동) 준수와 자원 효율성을 시에 완벽히 확보함.

6.2 한계점 및 향후 발전 방향

1. **통합 EDA 레시피 단일 모델 추론 미실시:** 문장 4유형 차등 증강, v5_reorder 프롬프트, CLIP 씬컷 힌트, 우도 K=4 TTA 등 개별 검증된 모든 EDA 결론을 하나로 통합한 단일 파인튜닝 모델의 최종 추론을 완료하지 못함.
2. **저자원 최고 효율 조합(4B + K=4 TTA)의 리더보드 미제출:** Holdout에서 +4.76%p 상승을 입증한 4B 모델과 우도 K=4 TTA의 조합을 최종 제출하지 못함.
3. **검증셋과 리더보드 간 분포 불일치 및 통계력 한계:** 252개 셔플 표본(holdout_300)의 통계력 한계로 인해, 검증셋 우위가 리더보드 점수 역전으로 이어지는 사례가 관측됨 (exp07 Holdout 48.41% ➔ Private 0.73983 하락, hard_perm Holdout +0.4%p ➔ Public -1.1%p 하락), 문장 유형 및 난이도 비율을 층화 반영한 확장 검증셋 설계가 필요함.
4. **약점 세그먼트의 직접적 체질 개선 미확인:** 최약점 유형(sparse_camX)에 대한 ×4 증강 배수 적용은 전체 점수를 상승시켰으나, 해당 세그먼트 자체의 정답률(20/83 유지)은 직접 보정되지 않음. 단순 사본 수 확대를 넘어 약점 유형 전용 학습법 도입이 필요함.
5. **조기 중단 시 저배수 유형 커버리지 유실:** ×2 증강에 머문 dense 계열이 조기 중단(977 Step) 시 완전 미학습 ID 유실의 81.8%를 점유함. 이를 해결하기 위해 [dense/sparse] × [camO/X] × [mark0/1+] 8유형 멀티티어 증강 및 Stratified DataLoader 도입이 필요함 (§4.4 참조).
6. **이미지 시공간 물리 힌트의 실시간 추론 활용 미완:** 물리/씬컷 정합률(~60%)의 한계로 인해 추론 파이프라인 탑재를 기각함. 신뢰도 향상 시 씬 분할을 통해 순열 탐색 공간을 24개에서 4개로 83.3% 축소하는 알고리즘의 실시간 소프트 가이딩 후속 연구가 필요함.
7. **비전 인코더(Vision Tower) 파인튜닝 미실시:** VRAM 8GB 자원 제약으로 인해 비전 인코더를 동결하고 LLM Attention 모듈에만 LoRA를 적용함. 유사 프레임 간 선후 혼동 오답을 근본적으로 극복하기 위해 비전 인코더 영역까지 LoRA를 확장하는 후속 실험이 필요함.