텍스트 단서와 이미지 프레임을 이용한 사건 순서 복원: 순열 증강과 균형형 다중 뷰 추론

Qwen3.6-27B의 RTX A6000 학습 및 단일 RTX 3090 추론 검증
Team 밥먹을돈으로3090사서거지됨 | 한지후(C), 박상률, 권성윤, 김범준 | github.com/SNUAICHAL/SNUAICHAL

Abstract. 본 연구는 문장과 무작위로 배열된 네 프레임으로부터 시간 순열을 복원하는 정확 일치 분류 문제를 다룬다. 모델 구조 자체의 등변성을 가정하지 않고, 입력 순열과 정답 좌표를 함께 변환하며 추론 결과를 동일한 원본 좌표계로 정규화하는 순열 좌표 일관 설계를 적용하였다. 테스트 시점 증강(test-time augmentation, TTA)에는 각 이미지가 네 표시 위치에 정확히 한 번 노출되는 네 개 순환 뷰(Latin4)를 사용하고, 정규화된 예측을 다수결로 집계하였다. Qwen3.6-27B 시각-언어 모델(vision-language model, VLM)에 4-bit 양자화 저랭크 미세조정(QLoRA)을 적용한 checkpoint-2726은 Public 정확도 0.93542를 기록하였다. 학습은 RTX A6000 48 GB 한 장에서, 최종 추론은 단일 RTX 3090 24 GB에서 수행하였다. 819행·3,276개 뷰에서 형식 오류는 0건이었고 추론의 NVIDIA Management Library(NVML) 기준 물리 피크는 21.15 GiB, 전체 시간은 3.89시간이었다. 동일 체크포인트의 생성 확률 집계는 0.92670, 24-view 다수결은 0.93193이었으며, step 3,400의 12-view 다수결은 0.93542로 동률이었다. 이에 같은 최고 점수에서 학습량과 추론 비용이 가장 작은 step 2,726·Latin4·다수결을 최종 구성으로 선정하였다. 최종 구성은 하나의 Qwen3.6 base와 하나의 LoRA adapter로 이루어지며, 같은 추론 절차로 외부평가 3,884행도 처리하였다.

그림 1. 실제 입력 메시지 형식과 최종 추론 과정. 학습에서는 이미지와 정답에 같은 순열을 적용한다.

1. 문제 정의 및 연구 목표

입력 표본은 문장 s와 이미지 $x=(x_1,x_2,x_3,x_4)$, 정답 y로 구성된다. S_4는 네 원소의 순열 24개 전체 집합이며 $y \in S_4$이다. 정답 숫자는 원본 프레임 번호가 아니라 현재 각 표시 위치 이미지의 시간축 위치를 뜻한다. 평가는 네 위치가 모두 일치해야 정답인 exact match이므로 두 장씩의 관계만 맞혀서는 부족하다.

순열 좌표 일관성. 본 방법은 등변 구조를 새로 설계한 것이 아니다. 학습 시 입력 순열과 정답 좌표를 함께 바꾸고, 추론 시 각 뷰의 출력을 원본 좌표계로 되돌려 집계함으로써 데이터 변환과 집계를 하나의 순열 작용으로 일치시켰다.

1.1 연구 문제

1. 입력 순열이 달라도 동일 사건에 대해 같은 원좌표 답을 낼 수 있는가?
2. 24개 전체 순열보다 훨씬 적은 뷰로 위치 노출을 완전히 균형화할 수 있는가?
3. RTX A6000에서 학습한 27B VLM을 단일 RTX 3090 24GB에서 재현 가능한 추론기로 실행할 수 있는가?

1.2 최종 시스템 구성

항목	고정값
Base	Qwen/Qwen3.6-27B, revision 6a9e13bd…e9
학습	운영진 제공 9,535행, primary stop 2,726
학습 장치	임대 RTX A6000 48 GB; peak 43.79 GiB
Adapter	QLoRA r/α=32/32, dropout 0.05, unmerged
추론 장치	단일 RTX 3090 24 GB; peak 21.15 GiB
추론 설정	NF4/BF16, SDPA, max area 512^2, batch 1
생성	64 tokens, thinking off, seed 42, strict S_4
TTA	1234/2341/3412/4123 → 원좌표 복원
집계	다수결(hard majority); 완전 동률=사전식

1.3 평가 근거의 구분

[A] Public 제출 결과. 전체 819행 중 70%인 573행으로 집계된 최종 또는 근접 후보의 추론 결과다. 나머지 30%의 Private 성능은 공개되지 않았다.

[B] 중복 억제 검증셋(clean954). 네 이미지의 byte SHA와 difference hash(dHash)가 학습 분할과 겹치지 않도록 구성한 954행의 Qwen3-VL-8B 개발 검증 결과이며, 최종 Qwen3.6의 holdout은 아니다.

[C] 수학·실행 검증. 위치 균형, 출력 형식, 메모리와 파일 동일성을 확인한 결과다.

1.4 오류 요인의 분해

겉으로는 24개 순열 중 하나를 고르는 문제지만, 오류의 원인은 세 가지로 나뉜다. 첫째는 장면의 실제 시간 관계를 잘못 읽는 내용 판단 오류, 둘째는 관계를 맞게 읽고도 현재 위치와 원래 위치 번호를 혼동하는 좌표 오류, 셋째는 설명문·중복 숫자·메모리 초과·불완전 파일처럼 정답 판정 전에 생기는 실행 오류다. 세 원인을 나눠 기록함으로써 모델의 판단 오류와 구현 오류를 구분했다. C_g는 순열 g로 배치한 뷰의 예측을 원본 위치 좌표로 되돌리는 역변환이며, 식은 3.1절에 제시한다.

실패 층	대표 위험	본 보고서의 검증
내용 판단	인접 프레임 모호성·문장 불일치	clean954·뷰 합의도
좌표	뷰 좌표를 그대로 투표	C_g 원좌표 정규화·왕복(round-trip) 검사
실행	형식 오류·메모리 부족(OOM)·불완전 CSV	S_4 검사·NVML·산출물 검사

1.5 후보 모델의 공통 검증 절차

모든 후보는 (i) 네 원본 위치의 일대일 대응, (ii) 원본 Id 순서, (iii) S_4 출력 범위, (iv) 모델·어댑터 동일성, (v) 단일 CUDA 작업을 먼저 확인했다. 이 조건을 만족하지 못한 결과는 점수가 높더라도 후보에서 제외했다. 반대로 실행 검증을 통과했다는 사실만으로 정확도가 높다고 볼 수 없으므로, 제출 결과·개발 실험·실행 검증을 구분해 표시했다.

2. 제안 방법

학습–추론 좌표 일관성. 입력 위치와 Answer를 함께 변환하고 모든 TTA 출력을 원본 위치 좌표로 정규화한다.

최소 완전균형 Latin4. 네 뷰만으로 image×displayed-position 노출행렬 N(i,j)=1을 달성한다.

대안 비교에 따른 선택. 생성 확률 집계, 24-view TTA와 추가 학습 step을 비교해 최종 구성을 단순화했다.

24 GB 추론 검증. 병합하지 않은 NF4 어댑터와 NVML 측정으로 모델 적재 전부터 종료 후까지 메모리를 기록했다.

파일 단위 재현성. 모델·어댑터·코드·입력·원시 출력·CSV의 SHA-256을 함께 기록했다.

방법론적 기여. 27B VLM에 TTA를 단순히 결합한 것이 아니라, 네 이미지의 순서를 하나의 좌표계로 보고 학습의 결정론적 순열 변환과 추론의 원좌표 정규화를 같은 규칙으로 구현했다. 뷰 수를 늘리기보다 좌표 일관성과 재현 가능한 집계를 먼저 확보했다.

2.1 주요 성능 및 실행 지표

지표	결과
최종 Public exact match	**0.93542**
final rows / views / parse fail	819 / 3,276 / 0
generation / end-to-end	약 3.75 h / 3.89 h
RTX 3090 physical peak	21.15 GiB / 24 GiB
A6000 학습	checkpoint-2,726 완료; physical peak 43.79 GiB
외부평가 실행	운영진 별도 무라벨 3,884행; 약 17.8 h
중복 억제 개발셋	clean954; train↔validation dHash/byte-SHA 0/0
학습 자료	9,535행; 38,140개 이미지

2.2 모델 선택 기준

- 실행 가능성 → 좌표·균형 검증 → Public 제출 결과 → 동점 시 적은 step/뷰 → 파일 재현성 순으로 판단한다.
- 서로 다른 모델·split·checkpoint가 함께 바뀐 비교를 단일 원인의 효과로 해석하지 않는다.
- 동일한 Public 정확도에서는 계산량과 재현 부담이 작은 구성을 우선한다.

2.3 비교 실험의 변인 통제

비교 실험은 (i) 동일 checkpoint·동일 4뷰에서 집계 방식만 변경한 경우, (ii) 동일 checkpoint·동일 집계에서 뷰 수만 변경한 경우, (iii) 학습 step과 뷰 수가 함께 달라진 경우로 구분했다. 앞의 두 비교는 각각 집계 방식과 뷰 수의 효과를 해석하는 데 사용했고, 복수 조건이 동시에 변한 결과는 후보 선별을 위한 관찰 비교로만 사용했다. 각 결과의 통제 수준과 해석 범위는 표 2에 함께 제시한다.

2.4 순열 좌표계의 일관성

index shuffle, QLoRA, TTA 각각은 알려진 구성요소다. 차별성은 이들을 한 방향의 순열 작용으로 묶은 연결 방식에 있다. 학습에서는 π가 입력과 Answer에 함께 작용하고, 추론에서는 g로 표시한 뷰의 출력에 원좌표 정규화 사상 C_g를 적용한 뒤에만 집계한다. 따라서 데이터 증강과 TTA가 서로 다른 좌표 규칙을 갖는 불일치가 생기지 않도록 했다.

일반 조합	본 시스템의 차이
shuffle만 적용	label 좌표까지 같은 π로 변환
여러 TTA를 직접 투표	각 뷰를 C_g로 원좌표 정규화한 뒤 투표
뷰 수를 늘려 탐색	노출행렬 하한을 만족하는 Latin4에서 정지
최고 점수만 기록	실패·비교 불가능성·파일 해시까지 함께 기록

2.5 부정적 결과와 설계 수정

clean954에서는 생성 확률 동률 처리가 근소하게 앞섰지만, 최종 Qwen3.6의 동일 원시 출력에서는 다수결이 더 높았다. 24-view TTA도 Latin4보다 계산량이 컸으나 정확도는 낮았다. 따라서 두 설정을 최종 구성에서 제외했다.

3. 순열 좌표 일관 데이터 변환

3.1 결정론적 순열 증강

항상 같은 위치에서 학습하면 사건의 내용보다 위치별 규칙을 외울 수 있다. 각 (seed, epoch, Id)를 SHA-256으로 해시해 $\pi \in S_4$를 정하고 입력 이미지와 정답을 동시에 변환한다. 동일 seed·epoch·Id는 같은 순열을 생성해 실행 재현성을 유지한다. 다만 이 방식이 24개 클래스를 정확히 균등 표집한다고 주장하지 않으며, shuffle 유무만 바꾼 완전 통제 실험도 수행하지 못했다.

원좌표 정규화. 뷰 순서를 $g=(g_1,\dots,g_4)$, 뷰 위치 예측을 $p=(p_1,\dots,p_4)$라 할 때, 원좌표 결과 $q=C_g(p)$는 각 j에 대해 q[g_j]=p[j]를 만족하도록 정의한다. 서로 다른 view 좌표의 예측을 직접 투표하지 않고, 모든 C_g(p)를 동일한 원본 위치 좌표계에서 집계한다.

3.2 데이터 분할 및 중복 통제

각 행은 Id, 네 이미지 경로(Input_1–Input_4), 사건 문장(Sentence)과 학습 정답(Answer)으로 구성된다. 제공된 학습 자료는 9,535행과 38,140개 이미지로 구성된다. 순서 정보가 불충분하다고 표시된 No_ordering 행은 1,478개(15.5%)였다. 문장 길이는 중앙값 26단어(IQR 15–32, 범위 5–69)였고, 이미지는 중앙값 640×360 px, 가로형 37,186개(97.5%)였다. 최종 Qwen3.6은 이 9,535행 전체를 학습했으므로 내부 holdout 정확도를 일반화 성능처럼 제시하지 않는다. clean954는 이미지 중복을 억제한 954행의 초기 Qwen3-VL-8B 개발 검증셋이며, random954는 같은 seed의 단순 무작위 954행 대조 분할이다. 두 분할은 최고 수치 탐색이 아니라 무작위 분할이 이미지 중복에 얼마나 민감한지를 확인하기 위한 대조로 사용했다.

split audit	train↔val overlap
clean954 dHash / byte-SHA	0 / 0
random954 dHash 영향 행	584 / 954
random954 byte-identical 행	458 / 954

같은 seed의 random954는 1,495개 difference hash(dHash)와 1,356개 바이트 해시가 학습·검증 분할을 가로질렀다. 따라서 높은 무작위 분할 수치를 모델 개선으로 오인하지 않고 최종 후보 선정에서 제외했다. 이 검사는 exact/근사 중복을 줄이지만 장면 독립을 완전히 증명하지는 않는다.

4. 모델 구조 및 학습 방법

Qwen3.6-27B[7]의 시각·언어 사전학습을 보존하면서 24 GB 추론 환경에서 실행하기 위해 4-bit QLoRA[1,2]를 적용했다. vision tower는 고정하고 language q/k/v/o 및 gate/up/down projection에 adapter를 학습했다. 사용한 revision은 2026년 4월 24일 공개본으로 대회 기준일인 2026년 5월 31일 이전이다. rank32의 정확도 우위는 통제 실험이 없으므로 주장하지 않고, 최종 모델의 설정으로만 기술한다. 학습은 RTX A6000 48 GB 한 장에서 수행했고 관측 최대 물리 메모리는 43.79 GiB였다. 최종 추론은 별도의 RTX 3090 24 GB에서 검증했다.

학습 항목	값
Optimizer	paged AdamW 8-bit
LR / schedule	1e-4 / cosine, warmup 0.03
Weight decay	0.01
Batch	device 1 × accumulation 8
Precision	NormalFloat 4-bit(NF4) base, bfloat16(BF16) compute
Seeds	seed/data_seed 42
Save / horizon	537 steps / 3 epochs
Training lineage	primary stop 2,726; post-hoc exact resume 2,726→3,400

4.1 학습 데이터 변환 절차

각 표본은 원본 이미지 목록과 정답을 먼저 파싱하고, sample key=SHA-256(seed‖epoch‖Id)에서 결정한 순열로 네 이미지 경로를 재배열한다. 정답은 재배열된 각 위치가 시간축 몇 번째인지 나타내도록 같은 사상으로 다시 쓴다. 입력 메시지는 재배열 뒤 생성하며, 증강 전후의 위치 동일성을 추론 기록에 남긴다. 이 순서를 뒤집으면 이미지와 정답이 서로 다른 좌표계를 갖게 된다.

단계	입력→출력	불변식
1 key	(seed,epoch,Id)→π	동일 run 재현
2 image	x→π·x	4개 경로의 일대일 대응
3 label	y→π·y	정렬값은 S_4
4 메시지	(π·x,π·y)→messages	이미지/정답 동좌표

4.2 체크포인트 선택

save interval 537은 복구와 학습 진행 관찰의 단위였다. 1차 학습은 step 2,726에서 종료했고, 이 지점을 최종 모델 후보로 동결했다. 이후 후기 학습의 실익을 확인하기 위해 동일 optimizer·scheduler·RNG 상태에서 step 2,726을 정확히 재개하여 674 step을 더 학습한 checkpoint-3,400을 별도 ablation으로 만들었다. checkpoint는 adapter뿐 아니라 optimizer, scheduler, RNG, trainer state를 함께 보존하고, 재개 시 global_step과 scheduler last_epoch를 대조했다. step 2,726에서는 512개 adapter tensor의 유한값과 archive SHA를 확인했다. step 3,400은 Public 0.93542로 동점이었지만 추가 학습과 12-view 추론을 요구해 최종 모델은 2,726으로 유지했다.

4.3 모델 규모와 QLoRA 설정의 해석 범위

27B 선택은 대규모 시각·언어 모델(VLM)의 사전학습 지식을 활용하려는 설계 결정이며 동일 split·동일 step의 8B/27B 통제 실험이 아니다. rank32 역시 최종 모델의 용량과 target module을 정의할 뿐 rank8보다 낫다는 근거가 없다. 따라서 '최종 모델에 사용했다'는 사실과 '다른 설정보다 우수함을 입증했다'는 주장을 구분한다. 이 구분은 코드 검증에서 재현되는 모델과 보고서의 성능 주장을 정확히 대응시키기 위한 것이다.

5. 최소 완전균형 Latin4 추론

5.1 Latin4의 균형 조건

Latin4는 각 이미지가 각 표시 위치에 정확히 한 번 노출되는 네 개의 순환 뷰 $G=\{1234,2341,3412,4123\}$이다. $N(i,j)=\Sigma_{g\in G}\,\mathbb{1}[g_j=i]$를 정의하면 모든 원본 이미지 i와 표시 위치 j에 대해 N(i,j)=1이다. 즉 네 뷰는 4×4의 1차 위치 노출을 정확히 한 번씩 채운다. 24개 전체 순열은 coverage를 늘리지만 이 1차 위치 편향은 이미 제거되며 forward pass만 6배가 된다.

view	pos1	pos2	pos3	pos4
1234	1	2	3	4
2341	2	3	4	1
3412	3	4	1	2
4123	4	1	2	3

수학적 범위. Latin4의 최소 균형은 위치 노출에 대한 성질이다. 임의의 불균형 4-view와 정확도를 직접 비교한 실험은 없으므로 성능 우위와는 구분한다.

5.2 원좌표 복원 및 예측 집계

각 뷰는 greedy generation 후 S_4의 순열 하나로 엄격히 해석한다. 설명문, 중복, 누락, 범위 밖 값은 형식 오류로 기록한다. 원좌표로 되돌린 예측의 최빈값을 선택하고 완전 동률은 사전식 최소값으로 처리한다. 생성 token의 생성 확률은 보정된 정답 확률이 아니므로 최종 집계에서 제외했다. 최종 3,276개 뷰의 형식 오류는 0건이었다.

6. 단일 RTX 3090 추론 최적화

base는 NormalFloat 4-bit(NF4), 계산은 bfloat16(BF16), attention은 scaled dot-product attention(SDPA), batch는 1, adapter는 merge하지 않은 상태로 유지한다. PyTorch allocator 값만 보지 않고 별도 process가 NVIDIA Management Library(NVML)의 전체 GPU 메모리 사용량을 0.5초 간격으로 model load 전부터 child 종료 후까지 측정했다. 최종 819행 추론의 최대 사용량은 22,710,861,824 B=21.15 GiB였다.

6.1 추론 파이프라인

1. CPU 사전 검사: CSV 열, 819개 고유 Id, 3,276개 참조 이미지, Answer 부재와 경로 범위를 확인.
2. model/adapter/source SHA 검증 후 단일 resident model을 load하고 Latin4를 순차 실행.
3. 행별 원시 출력·token log-prob·원좌표 예측·투표 결과를 감사(audit) 기록인 audit.jsonl에 저장.
4. submission.csv, audit.jsonl, metrics.json을 저장하고 SHA-256 및 NVML 측정 기록을 함께 보존.

6.2 최종 체크포인트 선정

checkpoint는 단순 저장 지점이 아니라 탐색 축이었다. step 3,400의 12-view TTA는 최종 모델과 Public 동점이지만 674 step과 3배의 뷰를 추가로 요구했다. '동점이면 이른 checkpoint와 적은 뷰'를 적용해 2,726을 선택했다. 두 요소가 함께 바뀐 비교이므로 이를 checkpoint 하나의 인과 효과로 해석하지 않는다.

검증 절차. 좌표 일관성, 최소 균형, 형식 검사와 24 GB 실행 가능성을 먼저 확인한 뒤 Public 결과를 비교했다. 이 순서는 정확도 변화와 구현 실패를 분리하기 위한 것이다.

6.3 Latin4의 최소성

한 view에서 원본 이미지 i는 표시 위치 하나에만 놓인다. i가 네 위치에 최소 한 번씩 노출되려면 view 수 k≥4가 필요하다. Latin4는 k=4에서 모든 i,j에 N(i,j)=1을 만족하므로 1차 위치 균형에 대한 하한을 정확히 달성한다. 이는 '네 뷰가 충분히 정확하다'는 통계 주장과 다르다. 증명한 것은 위치 노출의 최소 완전균형이며, 정확도는 별도의 Public 및 clean954 결과로 판단한다.

6.4 원좌표 정규화 예시

뷰 2341은 표시 위치 1,2,3,4에 원본 위치 2,3,4,1을 둔다. 이 뷰의 예측 $p=(p_1,p_2,p_3,p_4)$를 그대로 투표하면 다른 뷰와 숫자의 기준이 다르다. C_g는 $q_2=p_1$, $q_3=p_2$, $q_4=p_3$, $q_1=p_4$로 정규화한다. 모든 q가 원 slot 기준 S_4임을 확인한 뒤 count(q)를 계산하고, argmax 동률은 모든 행에 동일한 사전식 최소 규칙으로 해결한다.

구간	실측/계약	실패 시
pre-load	model tree·adapter·CSV 확인	CUDA 시작 전 중단
generation	4 views, 행별 진행(row-wise progress)	완전 prefix만 resume
aggregation	원좌표 S_4 다수결	형식 오류 기록
finalize	819 Id/order·CSV/audit	제출 파일로 사용하지 않음

6.5 추론 시간 및 메모리 분석

최종 실측은 16.4693초/row, 819행 generation 13,488초(약 3.75시간), end-to-end 13,990초(약 3.89시간)였다. TTA24는 93.8429초/row로 6배에 가까웠고 점수도 낮았다. final adapter는 약 637.6 MB로 55.6 GB base의 약 1.15%이며 merge하지 않아 원본 base와 adapter를 따로 검증할 수 있다. 최대 물리 피크(physical peak) 21.15 GiB는 24 GiB 카드에서 약 2.85 GiB의 관측 여유를 남겼다.

7. 실험 결과 및 분석

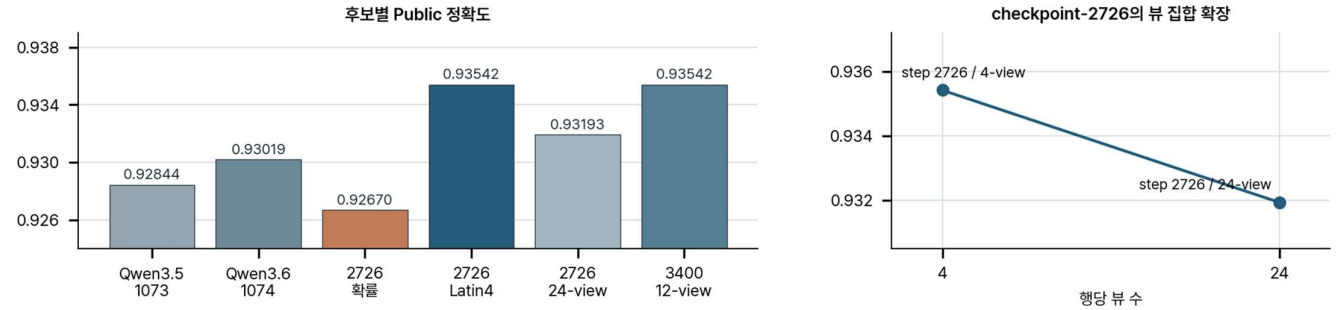

그림 2. 왼쪽: 최종 후보군의 Public 점수. 오른쪽: 동일 checkpoint-2,726에서 Latin4 네 뷰를 유지한 채 나머지 20개 순열을 추가한 4/24-view 뷰 집합 확장 비교.

표 1. 최종 후보의 Public 제출 결과. Private 점수는 모든 행에서 알 수 없으며, 함께 바뀐 조건을 구분해 해석한다.

모델/step	뷰	집계	Public	동일성·해석	결정
Qwen3.5-27B / 1,073	4	생성 확률	0.92844	이전 세대 대체안	대체
Qwen3.6-27B / 1,074	4	생성 확률	0.93019	초기 확인	참고
Qwen3.6-27B / 2,726	4	생성 확률	0.92670	같은 원시 출력; 다수결과 직접 비교	기각
Qwen3.6-27B / 2,726	**4**	**다수결**	**0.93542**	**최종; 사전식 동점 처리**	**채택**
Qwen3.6-27B / 2,726	24	다수결	0.93193	19,656개 뷰; 최종과 16행 상이	기각
Qwen3.6-27B / 3,400	12	다수결	0.93542	step·뷰 동시 변화; 동점 고비용	기각

표 2. 설계 선택별 근거와 해석 범위. A=Public 제출, B=clean954 개발 실험, C=수학·실행 검증.

#	검토 항목	비교 또는 관찰	결정과 해석 범위	근거
1	모델 선택	Qwen3.5-1,073 0.92844 → Qwen3.6-1,074 0.93019	탐색 근거; 모델·step 동시 변화	A
2	학습 종료 시점	1,074 생성 확률 0.93019, 2,726은 0.92670	학습량의 단조 개선은 확인되지 않음	A
3	집계 방식	2,726 동일 4뷰: 0.92670 → 0.93542	보정되지 않은 생성 확률 제거	A
4	TTA 뷰 수	2,726 다수결: 4뷰 0.93542, 24뷰 0.93193	균형 이후 추가 뷰의 잡음·비용	A
5	후기 checkpoint	3,400/12뷰 0.93542 동점	이른 step·1/3 뷰 선택	A
6	위치 노출 균형	Latin4에서 N(i,j)=1	정확도와 분리된 수학적 확인	C
7	좌표 정규화	뷰 좌표 직접 투표 금지; C_g로 원좌표 정규화	학습·추론에 같은 좌표 함수 적용	C
8	출력 형식	최종 0/3,276 형식 오류	설명문·중복·누락을 별도 처리	C
9	RTX 3090 실행	최대 물리 메모리 21.15 GiB	NF4+병합하지 않은 adapter	C
10	메모리 측정	allocator+분리된 로컬 NVML 0.5초 측정	driver·비-PyTorch 할당 포함	C
11	split 중복	clean954 partition 간 dHash/SHA 0/0	random split 오염 수치와 분리	B
12	index shuffle	결정론적 shuffle 구현; on/off 미실행	적용 사실만 기술, 향상 폭 미주장	C
13	LoRA rank	rank8 동일조건 대조 없음	최종 설정으로만 기술	C
14	단일 모델	base 1개+adapter 1개; 행별 혼합 없음	앙상블 없이 재현 부담 축소	C
15	외부 데이터	동일 설정으로 3,884행 형식 오류 0	실행 재현성만 주장; 점수 미지	C

주요 결과. 동일한 checkpoint-2,726·4뷰에서 다수결과 생성 확률 집계의 Public 차이는 5행이었고, 같은 checkpoint·다수결에서 24뷰와 4뷰의 차이는 2행이었다. 이를 강한 성능 우위로 해석하지 않고 작은 차이에서는 계산량과 구현 복잡도가 적은 구성을 선택했다. checkpoint-3,400 비교는 학습량과 뷰 수가 동시에 바뀌므로 후기 학습 하나의 효과로 해석하지 않았다.

표 3. 최종 후보의 계산량 비교. 상대 뷰 비용은 최종 Latin4=1.

후보	step	뷰/행	총 뷰	상대 비용	Public	최종 대비	결론
Qwen3.6 / 1,074 / Latin4 / 생성 확률	1,074	4	3,276	1×	0.93019	−0.00523	초기 확인
Qwen3.6 / 2,726 / Latin4 / 생성 확률	2,726	4	3,276	1×	0.92670	−0.00872	집계 기각
Qwen3.6 / 2,726 / Latin4 / 다수결	**2,726**	**4**	**3,276**	**1×**	**0.93542**	**0**	**최종**
Qwen3.6 / 2,726 / 24뷰 / 다수결	2,726	24	19,656	6×	0.93193	−0.00349	뷰 수 기각
Qwen3.6 / 3,400 / 12뷰 / 다수결	3,400	12	9,828	3×	0.93542	0	동점 고비용

7.1 비교 실험의 통제 수준

가장 강한 대조는 checkpoint-2726의 동일 3,276개 원시 출력에서 집계만 바꾼 생성 확률↔다수결이다. 그다음은 같은 checkpoint·다수결에서 뷰 수만 다른 Latin4↔24뷰 TTA다. 2,726/Latin4와 3,400/12뷰 TTA는 학습량과 뷰가 동시에 바뀌므로 '후기 학습이 무효'가 아니라 '추가 계산을 포함한 전체 구성이 개선되지 않았다'고만 해석한다. Qwen3.5↔Qwen3.6도 모델과 step이 달라 Qwen3.6을 시도한 근거일 뿐 모델 세대의 인과적 우위가 아니다. Public 분모는 819행의 70%인 573행이다. 집계 방식 비교의 +0.00872는 5행, 뷰 수 비교의 −0.00349는 2행에 해당한다. 24뷰는 Latin4와 16행의 예측이 달랐지만 순감소는 2행에 그쳤다. 행별 paired 결과가 없어 유의성 검정을 수행할 수 없으며, 이 차이만으로 통계적 우위를 주장하지 않는다. 따라서 다수결·Latin4는 성능 우위가 아니라 동등 성능 범위에서 계산량과 구현 복잡도를 최소화하는 근거로 채택했다.

대조	고정 조건	변경 조건	해석	해석하지 않은 범위
2,726 생성 확률↔다수결	model,step,4개 원시 출력	집계	동등권 단순 집계 선택	Private 보장
2,726 Latin4↔24뷰	model,step,다수결	뷰 집합/수	동등권 최소 뷰 선택	모든 24뷰 열등
2,726↔3,400	model,다수결 계열	step,뷰	계산량 포함 동점 비교	step 단독 효과
Qwen3.5↔Qwen3.6 초기 비교	task,4뷰/생성 확률	model,step	Qwen3.6 탐색 동기	Qwen3.6 일반 우위
clean954 Qwen3-VL-8B	split,model family 내	step,TTA	개발 방향성	최종 Qwen3.6 성능

7.2 Public 점수 기반 모델 선택의 한계

Public 점수는 573행에 대한 집계 정확도이므로 5행과 2행의 차이는 후보 순위를 안정적으로 구분하기에 작다. 또한 행별 paired 정오표가 없어 차이의 불확실성을 직접 계산할 수 없다. 따라서 이 범위에서는 점수 차이를 강한 우위로 해석하지 않고, 계산량과 분기가 적은 후보를 선택했다. 이 원칙은 반복 제출에 따른 선택 변동을 줄이고 최종 파이프라인의 재현성을 높이기 위한 것이다.

7.3 제한된 제출 예산에서의 비교 우선순위

예선 제출은 하루 2회로 제한되어 checkpoint별 성능 곡선과 동일 조건의 4/12/24-view 교차표를 모두 측정할 수 없었다. 제출 1회를 하나의 관측 예산으로 보고, 최종 선택을 바꿀 가능성과 변인 통제 수준이 큰 순서대로 동일 원시 출력의 집계, 동일 checkpoint의 뷰 수, 후기 checkpoint를 비교했다. 따라서 남은 공백은 step 또는 뷰 수의 독립 효과로 해석하지 않는다.

우선순위	비교	제출 1회당 정보	남은 공백
1	동일 원시 출력: 생성 확률↔다수결	집계 효과를 직접 분리	Private 성능
2	checkpoint-2,726: 4뷰↔24뷰	뷰 확대의 실익 확인	12뷰 단독 효과
3	2,726/4뷰↔3,400/12뷰	후기 전체 구성의 실익	step·뷰 독립 효과

8. Clean954 기반 개발 실험 및 오류 분석

clean954는 이미지 중복을 억제한 Qwen3-VL-8B 개발 검증셋으로, 최종 Qwen3.6의 holdout이 아니다. 이 실험의 최고 정확도는 246/954(25.79%)로, Qwen3-VL-8B Public 0.91972 및 최종 Qwen3.6 Public 0.93542와 차이가 크다. 이는 모델 규모·학습 범위·분할 난이도와 평가 분포가 모두 다른 결과다. random954에서는 584행에 dHash 중복 영향이 관찰되어 중복 기억이 무작위 분할을 부풀릴 가능성을 확인했다. 이 결과는 clean954와 random954의 분할 난이도 차이를 설명하는 범위에서만 사용했다.

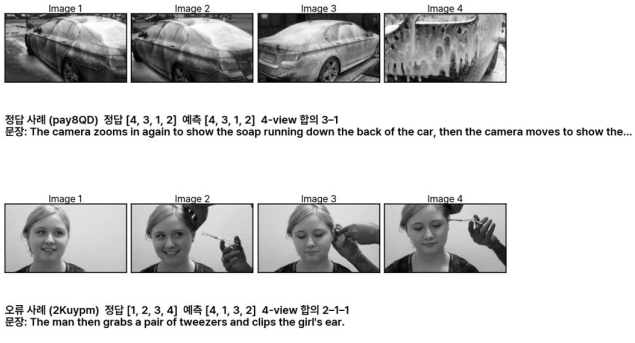

그림 3. 학습 자료의 정성 예시. 위: 순차적 상태 변화가 뚜렷한 성공 사례(Id=pay8QD). 아래: 짧은 문장과 미세 동작 변화가 결합된 실패 사례(Id=2Kuypm). 이미지·문장·예측을 함께 제시했다.

표 4. Qwen3-VL-8B clean954 실험.

step/precision	TTA	다수결	확률 동률	결론
3000 BF16	1	229	—	precision reference
3000 NF4	1	228	—	single-view best
3500 NF4	1	227	—	기각
3750 NF4	1	219	—	기각
4000 NF4	1	212	—	기각
4250 NF4	1	221	—	기각
4292 NF4	1	223	—	single-view
4292 NF4	**4**	**244**	**246**	**개발 기준**
3000 NF4	4	238	239	−7 vs baseline

8.1 합의도에 따른 정량적 오류 분석

clean954의 생성 확률 동률 처리는 246/954, 다수결은 244/954였다. 생성 확률 동률 처리가 15행을 고치고 13행을 틀려 순증가가 2행에 그쳤다. 최종 Qwen3.6에서는 방향이 역전되어, 생성 확률이 모델과 checkpoint를 넘어 보편적으로 유효한 집계 규칙이 아님을 확인했다.

agreement	rows	correct (conf-tiebreak)	accuracy
4	262	101	38.55%
3–1	308	79	25.65%
2–2	114	23	20.18%
2–1–1	216	38	17.59%
1–1–1–1	54	5	9.26%

예측 순열별 정확도의 최솟값 14.29%와 최댓값 38.89%는 각각 1/7과 7/18에서 계산되어 표본 수가 작다. 이는 정답 순열의 균형보다 모델이 특정 순열을 드물게 예측한 결과에 민감하므로 순열별 난이도의 근거로 사용하지 않았다. 외형 변화가 작은 연속 프레임, 반복 동작, 장면 전환, 문장–시각 불일치는 정답이 있는 개발 자료에서 관찰한 오류 유형이다.

8.2 Clean954 분할의 목적

무작위 행 분할은 같은 또는 거의 같은 이미지가 학습과 검증 양쪽에 들어가 시간 순서 자체보다 이미지 기억을 평가할 수 있다. clean954는 byte-SHA와 perceptual dHash component가 partition을 가로지르지 않도록 8,581/954로 분리했다. cross collision 0은 exact/근사 중복 통제의 근거지만 장면·촬영원·문장 수준의 semantic 독립까지 증명하지는 않는다. 따라서 생성 확률과 합의도의 경향을 최종 Qwen3.6 수치로 외삽하지 않았다.

8.3 뷰 합의도의 비지도 진단 활용

clean954에서는 만장일치 38.55%, 3–1 25.65%, 2–2 20.18%, 2–1–1 17.59%, 전부 상이 9.26%로 합의가 약할수록 정확도가 단조 감소했다. 이는 어떤 표본이 순열에 민감한지 나타내는 진단 지표이며, 감사 기록과 오류 분석 우선순위에 사용했다.

오류 유형	관찰 단서	현재 완화	남은 위험
미세 상태 변화	인접 뷰 불일치	문장+4 프레임 공동 입력	방향성 약함
반복 동작	2–2/2–1–1 투표	원좌표 다중 뷰	주기 대칭
장면 전환	한 프레임 이상치	전역 S_4 제약	카메라 단서 우세
문장–시각 불일치	문장과 변화 상충	27B VLM 사전학습	자료 모호성
좌표 민감성	순서별 답 변화	C_g+Latin4	인접 쌍 동시 배치 등 고차 편향

8.4 양자화 정밀도 비교

Qwen3-VL-8B step 3,000의 BF16/NF4 단일 뷰 결과는 229/228로 한 행 차이였다. 단일 checkpoint·split이므로 양자화 우위를 주장하지 않는다. NF4를 24 GB 후보로 유지할 기술적 근거로 사용했고, 최종 Qwen3.6은 물리 피크와 유한값 load를 별도 절차로 검증했다.

9. 제외한 대안 및 부정적 결과

생성 확률 동률 처리. 최종 동일 원시 출력에서 다수결보다 낮아 token likelihood를 정답 확률로 사용하지 않았다.

24-view TTA. 19,656뷰·93.84초/행에도 Latin4보다 낮아 추가 뷰를 기각했다.

후기 checkpoint. 3,400/12뷰는 0.93542 동점이었다. 674 step과 3배 뷰의 추가 계산을 기각했다.

무작위 분할. 584/954행 dHash, 458/954행 byte overlap. 점수 과대평가 가능성 때문에 제외했다.

전체 자료 학습. Qwen3.6이 9,535행 전체를 보았으므로 내부 평가 수치를 일반화 증거로 사용하지 않았다.

10. 최종 모델 선정과 일반화 고려

Public은 test의 70%만 반영하므로 반복 제출에 따른 선택 편향이 남는다. 따라서 최고 Public 점수가 같거나 차이가 2–5행에 그친 후보 중 학습량, 추론 횟수와 후처리 분기가 가장 적은 checkpoint-2,726/Latin4/다수결을 선택했다. 복잡한 후처리보다 동일한 원좌표 규칙과 단순 다수결을 유지해 미관측 분포에서의 변동과 구현 오류를 줄였다.

10.1 외부평가 재현성 검증

외부평가는 운영진이 별도 제공한 무라벨 데이터 3,884행에 대해 최종 모델로 한 차례 추론해 제출한 절차다. 추가 학습이나 모델 선택에는 사용하지 않았다.

항목	검증값
dataset	3,884행 / 15,536개 이미지
label boundary	Answer 없음; 추가 학습 0
model/adapter	final hash와 byte-identical
inference	Latin4 다수결, 형식 오류 0; 약 17.8 h(<24 h)
최대 GPU memory	25,403,994,112 B (23.66 GiB)
submission	한 차례, 3,884 unique Id
accuracy	운영진 미공개

운영 측 dataset refresh는 디렉터리 nesting과 check.txt만 바뀌었고 test CSV·sample submission·참조 이미지 tree의 내용은 동일했다. 완료된 2,911행을 Id 순서와 해시로 재확인한 뒤 나머지를 이어서 계산했다. 외부평가의 물리 피크 23.66 GiB는 본선 자료의 입력 분포와 장시간 실행이 달랐기 때문에 최종 시험 추론의 21.15 GiB보다 높았다. 차이를 단일 원인으로 단정하지 않으며, batch 1·NF4·단일 resident model로 24 GiB 이내를 유지했다.

11. 자원 효율성과 실행 조건

항목	적용 및 실측
학습	RTX A6000 48 GB; 최대 43.79 GiB; checkpoint-2,726 완료
최종 추론	RTX 3090 24 GB; 최대 21.15 GiB; 819행 3시간 53분 10초
외부평가	동일 RTX 3090 설정; 3,884행 약 17.8 h; 최대 23.66 GiB
모델 구성	Qwen3.6 base 1개+adapter 1개; adapter 미병합
저장 용량	base+adapter+tracked code 약 56.23 GB
학습 자료	9,535행; 38,140개 이미지

11.1 일반화 한계

- 최종 Qwen3.6의 중복 억제 동일-model holdout이 없다.
- rank8/rank32, image resolution, multi-seed, shuffle on/off의 완전 통제 대조가 없다.
- Latin4는 1차 위치 노출만 균형화하며 인접 쌍 동시 배치 같은 고차 순열 편향을 제거하지 않는다.
- Public 반복 관찰에 따른 selection overfitting과 Private 불확실성이 남는다.
- 외부평가 정확도는 공개되지 않아 실행 재현성 외 성능은 알 수 없다.

11.2 일반화 위험과 완화

위험	탐지/근거	완화	잔여 수준
Public 선택 편향	70% score 반복 관찰	동점 최소복잡도	중
domain shift	외부 label 미공개	동일한 단일 설정	중–고
분할 중복	dHash/SHA 검사	clean954 별도 분리	중
형식 오류	view별 출력 검사	S_4 엄격 검사; 최종 0	저
hardware variance	NVML 최대 사용량	batch1/NF4 여유	저–중
파일 변경	revision/tree/adapter SHA	manifest 검증	저

12. 재현성 검증

12.1 실행 절차

1. repository tag evaluation-final-v1 또는 고정 commit을 checkout하고 Python 3.10 환경을 설치한다.
2. scripts/download_weights.py로 29 files/15 shards의 base를 받고 manifest SHA를 검증한다.
3. scripts/download_final_adapter.py로 checkpoint-2726 adapter 두 파일을 받고 tensor 유한값을 확인한다. 이후 인터넷 연결은 필요하지 않다.
4. python -B -m scripts.run_final_inference --data-dir data --resume로 Latin4 다수결 추론을 실행한다.
5. scripts/validate_submission_artifacts.py로 Id 순서, S_4 Answer, 4-view 기록과 CSV 일치를 확인한다.
6. pytest와 Ruff를 실행하고 model/adapter/CSV SHA를 docs/final_results.json의 동결값과 대조한다.

12.2 소프트웨어 및 하드웨어 환경

구성	값
Python / Torch	3.10 / 2.6.0+cu124
Transformers / PEFT	5.13.1 / 0.19.1
bitsandbytes	0.48.2
precision	NF4; BF16 compute; SDPA
image	aspect preserved, max area 262,144 px
generation	64 tokens, thinking off, greedy
TTA/aggregation	Latin4 / 다수결 / 사전식 동률 처리
relative paths	data / models / weights / outputs

12.3 모델 및 산출물 무결성

대상	identity
model revision	6a9e13bd6fc8...f49c13e9
model tree	e4107e650879...a207ec07
adapter	189f6c1be09b...ca9f1f6
final CSV	75c4c223cd73...29b7563
source	inference/tta/submission/modeling SHA
external CSV	d75d2eeb4c4d...d48b17

model revision과 파일 목록, adapter, 실행 코드, 입력 자료, 원시 추론 기록과 최종 CSV의 SHA-256을 각각 기록했다. 이를 순서대로 대조하면 모델이나 코드가 달라진 최초 지점을 찾을 수 있다. 대회 데이터와 test prediction 원문은 GitHub에 포함하지 않고, Apache-2.0 기반 모델과 adapter 다운로드 manifest만 제공한다.

12.4 재현 실패 원인의 단계적 격리

재현 과정은 base revision과 tree, adapter, Python 환경과 source, 입력 Id·이미지, 행별 원시 출력, 원좌표 다수결, 최종 CSV의 순서로 동일성을 확인한다. 각 단계의 SHA-256과 행·뷰 수를 따로 기록하므로 최종 CSV가 달라질 때 최초로 달라진 계층을 찾을 수 있다. 형식 검사는 순열 범위뿐 아니라 원시 투표를 다시 집계한 결과와 CSV의 일치까지 확인한다.

12.5 재현 명령

단계	명령(요약)
base	python -B scripts/download_weights.py --manifest configs/weights/qwen36-27b-final.manifest.json --output models/Qwen3.6-27B
adapter	python -B scripts/download_final_adapter.py --output weights/qwen36-checkpoint2726
inference	python -B -m scripts.run_final_inference --data-dir data --resume
audit	python -B scripts/validate_submission_artifacts.py --test-csv data/test.csv --submission ... --audit ... --expected-tta 4
quality	python -m pytest; python -m ruff check src scripts tests

12.6 검증 스크립트의 범위

공개 검증 스크립트는 test Id의 비어 있지 않음·중복 없음·순서, submission의 Id/Answer 열과 S_4 형식, 행·뷰 수, 추론 기록과 CSV 정답의 일치, 형식 오류 수를 확인한다. 원좌표 복원과 다수결은 src/snuaichal/tta.py에서 C_g의 전 순열 왕복(round-trip) 검사와 다수결 단위 검사로 검증한다. 이 검사가 통과해도 정확도나 Private 성능이 증명되는 것은 아니다. base와 adapter의 크기 및 SHA는 별도 downloader에서 확인한다.

13. 논의

최종 구성은 운영진 제공 9,535행으로 학습한 단일 Qwen3.6-27B와 하나의 LoRA adapter를 사용한다. 학습의 순열 변환, 추론의 원좌표 복원, Latin4 다수결을 같은 좌표 규칙으로 묶었으며, 최종 산출물의 재현성은 실행 명령과 SHA-256 기록으로 확인할 수 있다.

13.1 방법론적 의의

- 문제의 순열 구조에서 데이터 변환·추론·집계를 일관되게 도출한다.
- 수학적 최소 균형과 Public·VRAM·형식 검증 결과를 구분해 해석한다.
- 부정적 결과까지 보존하여 최종 구성의 선택 논리와 계산 비용을 함께 제시한다.
- 외부평가에도 추가 조정 없이 동일한 단일 모델 추론 절차를 적용한다.
- 운영진이 원시 출력에서 CSV까지 독립적으로 다시 계산할 수 있다.

13.2 한계 및 후속 연구

가장 필요한 후속 실험은 (i) 최종 Qwen3.6의 중복 억제 다중 fold 평가, (ii) rank·해상도·shuffle의 matched ablation, (iii) Latin4가 제거하지 못하는 인접 쌍 동시 배치 등의 고차 순열 편향, (iv) 투표 합의도 보정, (v) 다중 seed와 domain shift 평가다. 후속 연구에서는 이들 조건에서 측정값의 분산과 일반화 성능을 직접 측정할 필요가 있다.

13.3 기존 기법과의 차이

Deep Sets[3]와 Set Transformer[4]는 집합 원소의 순서를 제거한 permutation-invariant 함수를 다룬다. 반면 본 과제는 순서 자체가 출력이므로 입력 위치 정보를 버릴 수 없다. 본 연구의 차별성은 구조적 등변 모델을 새로 제안하는 대신, 입력 순열을 좌표 작용으로 정의하고 학습 정답 변환과 TTA 원좌표 정규화를 같은 함수로 구현한 점에 있다. 또한 TTA 집계가 기존의 정답을 오답으로 바꿀 수 있다는 보고[5]와 본 실험의 생성 확률/다수결 결과를 함께 고려해, 뷰 수 증가를 자동적인 개선으로 간주하지 않았다. 일반적인 이미지 증강 분류와 설계 쟁점은 관련 survey[6]를 참고하였다.

기존 구성요소	추가한 설계	검증 자료
Qwen3.6-27B	입력 위치를 명시적 좌표계로 처리	Qwen3.5/Qwen3.6 비교와 해석 범위
QLoRA	24 GB 단일 GPU용 병합하지 않은 adapter	로컬 NVML 21.15 GiB
TTA	위치 노출 하한을 만족하는 네 뷰	N(i,j)=1 Latin4 증명
다수결	각 뷰를 원래 좌표로 정규화한 뒤 집계	C_g 전 순열 왕복(round-trip) 검사
비교 실험	채택·기각 결과와 해석 범위를 함께 기록	15개 설계 선택 기록

13.4 설계 결정의 근거

설계 항목	선정 근거
TTA 뷰 수	균형 하한을 만족한 뒤 뷰를 늘려도 점수가 개선되지 않았다.
집계 방식	동일 원시 출력에서 다수결이 생성 확률 집계보다 높았다.
체크포인트	checkpoint-3400은 동점이지만 674 step과 3배의 추론 뷰가 추가되었다.
일반화 고려	동점 후보 중 단일 모델·최소 계산 구성을 택해 선택 자유도를 제한하였다.
재현성 진단	tag, manifest, adapter, 추론 기록, CSV 순으로 동일성을 검증한다.

14. 결론

본 연구는 무작위로 제시되는 네 이미지의 순서를 하나의 좌표계로 해석하고, 학습 시 순열 변환과 추론 시 원좌표 복원을 일관되게 적용했다. 각 이미지가 네 표시 위치에 한 번씩 나타나는 최소 완전균형 Latin4와 다수결은 24-view 추론보다 적은 계산으로 더 높은 Public 점수를 얻었다. checkpoint-3,400은 같은 점수를 기록했지만 학습 step과 추론 횟수가 늘어 checkpoint-2,726을 최종 선택했다. 최종 모델은 Public 0.93542, 형식 오류 0건, RTX 3090 최대 21.15 GiB를 기록했으며 같은 설정으로 외부평가 3,884행 추론도 완료했다. 이 결과는 모델 크기뿐 아니라 입력·출력 좌표의 일관성, 최소 균형 추론과 비교 가능한 실험 설계가 중요함을 보여 준다.

참고문헌

[1] T. Dettmers et al., "QLoRA: Efficient Finetuning of Quantized LLMs," NeurIPS, 2023.
[2] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR, 2022.
[3] M. Zaheer et al., "Deep Sets," NeurIPS, 2017.
[4] J. Lee et al., "Set Transformer," ICML, 2019.
[5] D. Shanmugam et al., "Better Aggregation in Test-Time Augmentation," ICCV, 2021.
[6] C. Shorten and T. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," J. Big Data, 2019.
[7] Qwen Team, "Qwen3.6-27B model card," revision 6a9e13bd...f49c13e9, 2026.

실험 및 공개 범위

학습·추론에는 운영진 제공 자료와 단일 모델을 사용했다. 외부 상용 API·외부 학습 자료·의사라벨·모델 앙상블·테스트 정답 수작업은 사용하지 않았으며 API 비용은 0원이다. 성능 표에는 직접 생성하고 해시를 보존한 결과만 포함했고, 공개되지 않은 Private·외부평가 정확도는 기재하지 않았다. 저장소에는 대회 원자료와 예측 원문을 포함하지 않는다.