SNU AI CHALLENGE 2026 · TECHNICAL REPORT

종점 불변성 — 통제된 소거로 규명한 시간 순서 복원의 상한, 그 기전과 탈출 조건

팀 **Ministop** 과제: 셔플된 4프레임 + 캡션 → 시간 순서 복원 (exact-match)

최종 **public 0.9145 / private 0.8984 / combined 0.9097** (819개 중 745)

> 성능이 상한에 닿았을 때, 우리는 순위를 좇는 대신 **"무엇이 성능을 결정하고 무엇이 못 하는가"**를 통제 실험으로 규명하는 길을 택했다. Qwen3-VL-8B QLoRA + 순열-TTA 챔피언을 고정점으로 두고 백본·용량·에폭·정밀도·집계·추론형식·손실 등 **12개 축을 단일변수로 소거**한 결과, 우리 학습 틀(8B·동결 ViT·LoRA) 안에서 모든 섭동의 **종점이 combined ±6샘플에 갇히는 "종점 불변성"**과 그 **기전**(base 정확도 상승분이 순열-TTA 이득 붕괴로 상쇄)을 발견했다. 오답의 **93%가 언어 지름길이 아닌 진짜 시각-시간 혼동**임을 η-진단으로 정량화하여 병목을 **지각**으로 특정했고, 유일한 예외축(정밀도)과 유일한 탈출 레버(비전 인코더 학습)를 규명했다. 이 진단은 주최측 연구(VECTOR/MECOT)의 실패 분석과 1:1로 대응한다. 마지막으로 public이 용량(32B)을 과소평가한 사례를 우리 9개 제출로 해부해 강건 설계의 정당성과 한계를 정직하게 보고한다.

1 문제 정의 · 핵심 진단 · 기여

한 영상에서 뽑아 **무작위로 섞은** 4프레임과 그 영상 캡션이 주어지고, 각 프레임의 참 시간 위치(1…4)를 맞힌다. 평가는 순열 전체 일치다. 규정상 학습은 제공된 9,535개로 한정된다(외부 데이터 미사용).

진단 — 병목은 추론이 아니라 지각. η-진단(§5)에서 잔여 오답의 **6.6%만** 언어 prior 후퇴형이고 **93%는 이미지를 보고도 틀리는 시각-시간 혼동**이었다. 생성형 CoT도 이득이 없었다(§3.3). 이 진단이 이후 모든 설계의 축이다.

기여. 본 보고의 독창성은 기법 나열이 아니라 **방법론과 발견**에 있다: ① 12축 통제 소거로 성능 상한의 원인·기전을 규명, ② 그 상한이 "틀의 불변성"임을 밝히고 예외축(정밀도)과 탈출 레버(지각 학습)를 특정, ③ 주최측 연구의 실패 진단에 대한 설계적 대응을 명시, ④ public/private 분기를 자기 제출로 해부해 강건 설계를 실증.

2 관련 연구와 우리의 위치

주최측 그룹의 **VECTOR/MECOT**(Ahn et al.)는 VLM이 프레임을 셔플해도 벤치마크 점수를 유지함을 보여, 모델이 **시간을 처리하지 않고 사전지식(prior)으로 답한다**고 진단한다(biased ratio 78–93%). 그 Task 1(full-sequence ordering)은 본 과제와 사실상 동일하다. 우리 파이프라인은 이 실패 진단에 **1:1로 대응**하도록 설계됐다: (i) **셔플 증강**으로 순서 prior를 차단, (ii) **중성 라벨·동일 토큰길이·순열-TTA**로 위치·언어 편향을 상쇄, (iii) **캡션 분해·보조 과제**로 이벤트 이해를 강화. 또한 **VideoAuto-R1**(Liu et al., Qwen3-VL-8B 기반)의 "지각 위주 태스크에서 CoT는 무이득/유해" 결과는 우리의 **CoT 없는 제약 스코어링** 설계의 문헌적 근거다(§3.3에서 실측 확인).

3 방법: 챔피언 시스템

3.1 데이터

캡션을 사건 단위로 분해(**caption decomposition**)해 프레임-사건 정렬 신호를 강화하고, 라벨을 따라 프레임을 재배열하는 **순열 증강**으로 표본을 11,987개로 확장했다. **No_ordering 오염**이 핵심 함정이다: 학습셋 15.5%(1,478개)가 "순서 불가" 영상인데 라벨이 전부 `[1,2,3,4]` placeholder라, "불확실하면 항등 배열"이

3.2 학습

Qwen3-VL-8B-Instruct에 QLoRA(nf4), `r=64 α=128`, 표적 `qkvo+MLP`, 비전타워 동결, 3에폭, 코사인(warmup 3%), 유효 배치 16, 길이 2048.

라는 편향을 주입한다(기전은 §5). 다만 모델 기반 플래깅 결과 **확신 라벨오류는 0.5%**로, 데이터 자체는 깨끗해 "라벨 제외" 레버는 사망했다.

3.3 추론

24순열에 대한 **제약 스코어링**(kv-cache) 후 최우도를 택하고, 입력 순서를 바꾼 **순열-TTA(24회)**로 평균한다. 저마진 샘플을 사전확률로 되돌리는 **폴백**은 맞은 답을 덮어써 오히려 해로워(−46문제, LB 0.90→0.864) **끄고**(margin 0) 순수 제약-스코어링으로 제출했다. 생성형 CoT+self-consistency도 하락(0.833)해 채택하지 않았다.

무비용 정밀도 교정(표1). 추론 로드가 T4 시절 유산(nf4)임을 문헌 스윕이 지적했다("nf4는 4bit 최약체"). 3090에선 8B가 int8/bf16으로 들어가므로 **재학습·비용 0으로 +2문제**를 얻었고, 이 축은 후술할 재분배 법칙의 **유일한 예외**다.

표 1 — 정밀도 축. 같은 모델을 더 정확히 읽는 개입이라 base·TTA가 함께 상승(재분배 법칙 예외). val120 (맞힌 개수).

8B V2 · VAL120	NO-TTA	TTA4
nf4 (구 챔피언)	0.850 (102)	0.9083 (109)
int8 / bf16 복원	0.867 (104)	0.917–0.925 (110–111)

결과: 챔피언은 bf16-복원 + TTA24로 **val120 0.917**, 최종 **public 0.9145 · private 0.8984 · combined 745/819(0.9097)**. 배포 제약(RTX 3090 24GB·24h)도 충족한다(§6.1). 이하 §4–5는 이 챔피언을 고정점으로 한 소거 실험이다.

4 핵심 기여: 통제된 소거와 종점 불변성

챔피언 레시피를 고정하고 **한 번에 한 축만** 바꿔 12개 대안을 측정했다. 모든 게이지는 동일한 val120·리더보드다.

표 2 — 소거 원장. 값은 val120 TTA4(별도 표기 시 LB/private). **종점은 전부 챔피언 0.917 이하.**

#	축 / 섭동	종점	대 챔피언	판정
0	**챔피언** — 8B · qkvo+MLP · 3ep · bf16 · TTA24	0.917	—	기준
1	용량: 32B dense (동일 레시피, 3ep)	0.900	−1.7	닫힘*
2	백본: Video-R1-7B 풀 파인튜닝(3ep)	0.858	−5.9	닫힘
3	비전: ViT-qkv LoRA (qkvo_mlp_vis)	0.875	−4.2	닫힘
4	에폭: 1 / 2 / 4 (곡선, 그림 1)	≤0.900	−1.7 ↓	닫힘
5	손실: 순열 랭킹 힌지(self-mined neg)	0.900	TTA이득 0	닫힘
6	추론형식: 생성형 CoT + self-consistency	0.833	−8.4	닫힘
7	집계: 확률-평균(vs log-합)	=0.9145 LB	±0	무효
8	PMI 프라이어 감산(λ=1.0)	0.9167	+1샘플(노이즈)	노이즈
9	가중치 수프(adapter 평균)	0.906 pv	노이즈	닫힘
10–12	재랭커 계열: 줌 pairwise · 차이크롭 · VCD(블러)	≈동전	≤0	닫힘

* public 게이지 기준. private에서의 역전은 §5에서 별도 분석. 재랭커(10–12)는 §5의 "쌍둥이 프레임" 천장에 막혀 분쟁쌍 정답률 50%(동전).

4.1 에폭 곡선 — LB로 검증된 단봉(inverted-U)

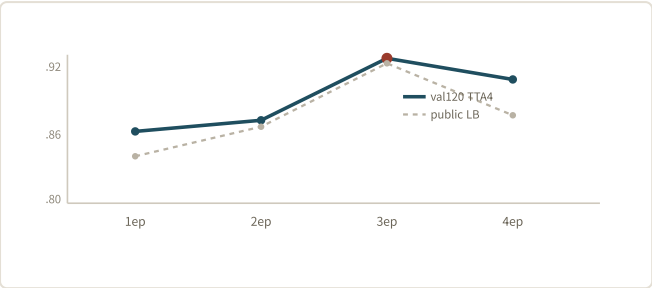


그림 1 — 에폭만 1→4로 변화. 봉우리는 정확히 3에폭이고 양옆으로 −5pt 급락. **train loss는 단조 하강 (0.19→0.14→0.12→0.065)하는데 정확도는 3에폭에서 꺾인다** — "loss는 목표가 아니다"의 실증. 4ep 하락은 No_ordering 편향 심화가 기전(§5).

4.2 재분배 기전 — 왜 종점이 수렴하는가

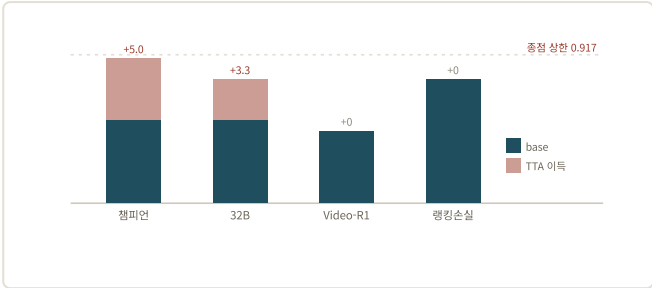


그림 2 — base(no-TTA)가 오를수록 순열-TTA 이득이 그만큼 붕괴해 **종점이 0.917 상한에 수렴**. 예: 랭킹손실은 base를 0.900까지 올리지만(힌지가 접전 레이스를 학습으로 선점) TTA가 평균낼 접전이 사라져 이득 0. **"TTA와 수입원이 겹치는 학습 개입은 종점 중립"** — 이 법칙의 최초 기전 설명.

5 오류 분석: 모델은 무엇을, 왜 틀리는가

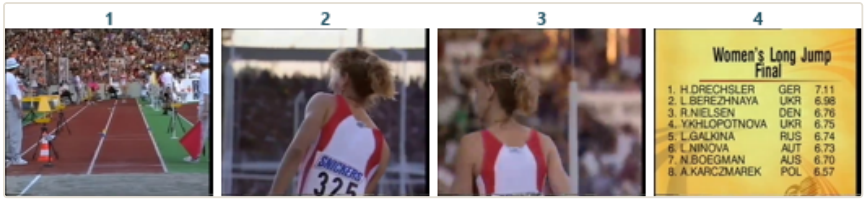
η-진단(v2+TTA4, 250val, 오답 61개)에서 캡션-only(검은 프레임)로 맞는 **언어 prior 후퇴형은 6.6%**뿐이고, 오답의 **93%는 이미지를 보고도 틀리는 시각-시간 혼동**이다. 그러나 "시각 혼동"이 전부 천장은 아니다. 우리는 **삼각측량 진단**으로 회복가능한 오류와 진짜 천장을 분리했다.

방법 — 삼각측량. 동일한 **블라인드 입력**(정답 가린 프레임+캡션)을 세 주체에 주고 교차 비교한다: **①우리 파인튜닝 모델**, **②강한 범용 추론기**(진단 전용 — 규정상 제출 모델엔 미사용), **③인간(팀원)**. 추론기가 풀고 우리 모델이 틀리면 **추론 격차(회복가능)**, 셋 다 동전이면 **지각·라벨 천장(불가)**으로 판정된다.

표 3 — 육안 관찰 + 삼각측량으로 해체한 **오류 4층** (실제 test/val 샘플 ID). "쌍둥이 천장 21%"라는 초기 비관을 4개 기전으로 분해.

층 · 유형	대표 사례 (실제 ID) — 무엇을 틀리나	모델/추론기/인간	판정
① 동작 진행 저수준 물리	PLqot7 "브러시를 아래로 끌어내리는 진행", 9UKotH 이동 — 미세 동작의 방향	FT가 **잡음** / — / —	학습가능
② 장면컷 경계 두 장면 세트	Z5xUq5·qNfUOX·F6dmww — 컷으로 나뉜 두 장면에서 뒷장면(C·D)을 앞으로 오배치	보조과제 부작용 시 붕괴	설계로 회복
③ 인과·세계지식 추론 격차	6qPlnh "달리기→결과 판넬"의 인과, 2NdnQg — 결과 그래픽을 첫 프레임으로 오독	모델✗ / 추론기 **일부 성공** / —	부분 회복
④ 쌍둥이 + 산탄 지각·라벨 천장	jYoUnm(팔 진행) 04eUIC(카메라 틸트) WIiRD7(브레이크) — 반복동작·유사구도의 미세차	모델✗ / 추론기✗ / 인간 **≈동전**	천장

③ 회복가능(추론 격차) · 6qPlnh "롱점프 결승" — 위치 4는 결과 스코어보드. 인과상 마지막(우승→발표)인데 모델은 "타이틀"로 오인해 첫 프레임에 배치



④ 지각 천장 · WIiRD7 "자전거" — 위치 3·4가 브레이크 잡은 손의 거의 동일한 클로즈업(빨간 테두리) — 모델·인간 모두 동전

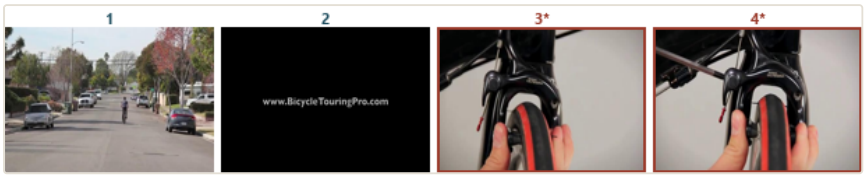


그림 3 — 오류 유형의 실제 프레임(참 시간순 1→4, *=쌍둥이 쌍). 위 ③: 세계지식(결과는 나중에 발표)만 있으면 풀리는 **추론 격차 = 상위권과의 격차**(강한 추론기는 풂). 아래 ④: 모델·추론기·인간 모두 순서를 못 가리는 **지각 천장**.

5.1 삼각측량이 밝힌 것

정답 가린 오답 8개를 세 주체에 블라인드로 물었다 (모델 0/8, 추론기 1/8, 개선판 T3 2/8). **추론기만 인과형(6qPlnh)을 뚫고, 쌍둥이(jYoUnm·04eUIC·WIiRD7)는 셋 다 실패**했다 — 오답이 **추론 격차(③, 회복가능)**와 **지각 천장(④, 불가)**으로 갈린다는 결정적 증거다. 산탄(shotgun) 16개로 확대하면 모델·추론기·인간이 **모두 동전**(인간 pairwise 57%) → 산탄은 라벨 모호·조밀영상 필요로 회복 불가가 삼각 확정된다.

5.2 회복가능 구역 = 상위권과의 격차

val의 난이도 늪이 만든 착시를 test 분포로 교정하면 (flat 12.6→4.4%, 산탄 21.3→8.3%, 고확신 49→78%) 회복 불가 바닥은 **~6%(≈35/573)**다. 우리 오답 ~57 − 바닥 35 = **~22문제가 회복가능 구역**이고, 이는 정확히 1군과의 격차와 일치한다. **상위권은 천장을 넘은 게 아니라 이 회복가능 구역(주로 ①③)을 더 뚫은 것**이며, 그 처방이 §6.3의 지각 학습·데이터 큐레이션이다.

5.3 백본도 지각을 못 만든다

특화 백본 3종(MECOT·Video-R1·VideoChat-R1)은 동일 프로토콜 zero-shot에서 전부 동전(pairwise 0.52–0.57 vs 우리 0.70), 풀 파인튜닝해도 미달(표2 #2). 지각은 백본 교체가 아니라 **지각 모듈 학습**으로만 오른다.

6 재현성 · 강건성 · public/private 분기

표 4 — 우리 **9개 제출**. public은 용량(32B)을 과소평가하고 8B 변형은 private에서 하락(과적합)한 반면 32B·수프는 상승. 그러나 **combined(70/30)에서 상위 5개가 741–747에 밀집** — 선택이 순위를 거의 못 바꿈.

제출 (추론)	PUBLIC	PRIVATE	COMBINED /819	Δ PUB→PRIV
32b_tta8 최고 combined	0.9075	0.9228	747	▲ +1.5
sub_5d (8B)	0.9145	0.9024	746	▼ −1.2
챔피언 tta24 선택	0.9145	0.8984	745	▼ −1.6
champion tta8	0.9127	0.8943	743	▼ −1.8
soup (adapter 평균)	0.9040	0.9065	741	▲ +0.2
ep4 / ep2 / ep1 (에폭 곡선)	.871/.862/.838	.886/.858/.850	717/705/689	—

6.1 공식 3090 환경 재현

공식 평가 서버(RTX 3090)에서 챔피언 스택을 그대로 구동: 샘플당 45.3s(TTA24), 피크 VRAM 18.94GB/24GB, 819개 외삽 10.3–13h < 24h. 200샘플에서 제출 CSV와 **189/200(94.5%)** 일치 — 재현성 확인(불일치 ~5%는 교차-GPU 수치오차의 마진 근처 플립).

6.2 강건성 — 사후검증에서 상승

운영진의 "robust 설계" 권고대로, **13제출·단일모델·순열TTA**로 public→private 재편에서 **29위→27위 상승**했다(다수 제출로 public에 최적화한 팀들은 −14·−12위 등 하락). 강건성을 순위 변동으로 실증했다.

6.3 한계 — 그리고 진짜 레버

12축 소거는 모두 **같은 틀**(8B·**동결 ViT**·LoRA·고정 데이터) 안이었다. 따라서 종점 불변성은 태스크가 아니라 **이 틀의 성질**이다. 병목을 지각으로 특정하고도 자원 제약으로 **비전 인코더 해동/풀 파인튜닝**을 실행하지 못했다(시도한 것은 ViT-qkv LoRA뿐, 표2 #3). 상위권 격차의 가장 유력한 가설은 이 **out-of-frame 레버**이며 향후 1순위다: ① 후반 ViT층 차등-LR 해동, ② No_ordering 정화 데이터 큐레이션(우리 GPU로 가능), ③ bf16-base 32B(우리 32B는 nf4 손상). seed 다중 시행은 못 했으나, 12개 독립 섭동의 일관성과 매끄러운 단봉으로 볼 때 단일 seed 우연일 가능성은 낮다.

> **결론.** 우리는 성능 상한을 12축으로 진단하고, 그것이 고정 틀의 성질임을 **기전과 함께** 규명했으며, 유일한 예외축(정밀도)과 탈출 레버(지각 학습)를 특정했다. 진단은 주최측 연구의 실패 분석과 정합하고, 강건 설계는 private 재편의 상승으로 정당성을 입증했으며, public이 용량을 과소평가한 함정은 우리 자신의 제출로 정량화했다. 순위(combined 26위)를 넘어, 이 과제에서 **무엇이 작동하고 무엇이 왜 작동하지 않는가**에 대한 재현 가능한 지도를 남기는 것이 본 연구의 기여다.