캡션 기반 셔플 프레임의 시간 순서 복원을 위한 다중모달 순열 스코어링과 선택적 재순위화

Team AIKU(노윤헌, 마현우, 장서현)

2026년 7월 28일

요 약

본 연구는 무작위 순서로 주어진 네 장의 프레임과, 단일 캡션을 입력으로 받아 올바른 순서에 해당하는 순열을 반환해야 하는 문제에 대해 다룬다. Qwen3-VL-8B-Instruct 모델을 Backbone으로 사용하였으며, 가능한 24가지의 순열 후보에 대하여 teacher-forcing 방식으로 VLM이 가장 높은 log-probability를 보이는 후보를 정답으로 채택하는 방식을 사용하였다. 주어진 데이터셋을 이용해 VLM을 Lora Fine-Tuning 하였으며, 학습 과정에서는 Frame shuffle을 통한 Data Augmentation 기법, p(Caption | GT Frame)을 맞추는 문제를 보조 QA Task 1으로 학습, Frame A와 B의 시간 순서를 맞추는 문제를 보조 QA Task 2로 학습하는 등의 방법론을 사용하였다. 추론 과정에서는 KV 캐싱을 통한 추론 가속화, Test Time Augmentation을 통한 보정, 시각 구조 보정, p(Caption | GT Frame) Task Score를 통한 Reranking 기법을 결합하였다. 그 결과 zero-shot 공개 리더보드 0.710에서 최종 0.93193이라는 점수를 달성하였다.

1 Introduction

본 과제의 입력은 영상을 설명하는 캡션과 무작위 순서의 프레임 네 장이며, 출력은 원래의 시간 순서를 나타내는 순열이다. 모델이 해당 과제를 올바르게 수행하기 위해서는 자연어 캡션의 의미에 대한 이해, 이미지-자연어 의미 정렬, 그리고 이미지의 시간 순서에 대한 이해가 동반되어야 한다.

이러한 과제는 MLLM 계열 연구에서 지금까지 다루어진 사례가 있는 영역이다. 우리 팀은 먼저 철저한 사전 조사가 필요하다고 생각하고, 관련 연구에 대해 리서치를 진행하였다.

먼저 TempVS 벤치마크를 낸 Paper [9]에서는 멀티모달 모델이 이미지 시퀀스의 시간 구조를 이해하는지 평가하는 벤치마크를 제안했고, TimeBlind [10]는 최신 Video LLM들이 프레임을 순서 없는 집합처럼 취급하는 'bag-of-frames' 문제를 정량화하였다. 또한 TPRU [11]에서는 3-4장의 셔플된 프레임과 텍스트 설명을 주고 올바른 순열을 출력하도록 하는, 이 과제와 굉장히 비슷한 Task에 대하여 7B급 모델의 타깃 SFT만으로 GPT-4o를 추월함을 보였다. 우리는 이들 선행 연구에서 다뤄진 후보 모델들을 동일 검증 표본에서 직접 비교하여, 성능이 더 좋은 모델을 Baseline으로 채택하였다(3.2절).

이미지의 시간 순서에 대한 이해가 필요한 만큼, 비디오 모델을 Backbone으로 사용하는 것도 고려하였으나, 추론 시간·메모리 제약(RTX 3090 24GB, 24시간)을 감안해 검증된 8B급 VLM 계열을 우선하였다.

핵심 기여는 세 가지이다. 첫째, 이미지와 캡션을 공통 prefix로 처리한 뒤 24개 후보의 조건부 로그확률을 비교하여 형식 오류를 제거하고 top-1과 top-2의 점수 차인 마진을 신뢰도 지표로 얻었다. 둘째, 순방향 점수 p(순서 | 캡션, 프레임)와 역방향 점수 p(캡션 | 정렬된 프레임)를 결합하고, 역방향 능력을 캡션 생성 과제로 학습하였다. 셋째, 어려운 데이터층의 oversampling이 일반화가 아닌 암기를 유발함을 자기-EM 진단으로 확인하고 최종 학습 구성을 수정하였다.

2 Related Works

2.1 프레임 순서 예측과 형식화

프레임 순서 예측은 이진 순서 검증(Shuffle & Learn [6]), 4프레임 정렬(OPN [7]), $n!$-way softmax 분류(VCOP [8])로 이어지는 자기지도 학습 계보에서 다뤄져 왔다. n=4처럼 후보 공간이 작은 경우 전체 순열을 전수 평가할 수 있다는 관찰이 본 연구의 착안점이다.

2.2 TempVS

TempVS [9](ACL 2025 Findings)는 멀티모달 모델이 이미지 시퀀스의 시간 구조를 이해하는지 평가하는 벤치마크로, 이벤트 문장이 조건으로 주어진 이미지 순서 판정 과제를 포함해 본 과제와 형태가 가장 가깝다. 상용 최강 모델도 이 과제 유형에서 매우 약했고(GPT-4o 23.0%), 오픈 모델 중 InternVL-MPO 계열이 상대 강세였다. 단 이는 구세대(2024–2025년 초) 모델 기준의 결과로, 본 연구에서는 이를 참고하되 후보 모델들을 자체 파일럿에서 직접 비교해 Baseline을 선정하였다.

2.3 TimeBlind와 신세대 VLM의 시간 추론 한계

TimeBlind [10](2026)는 모델이 프레임을 순서 없는 집합처럼 취급하는 'bag-of-frames' 문제를 정량화하고, 최신 모델에서도 이 결함이 지속됨을 보였다(order sensitivity 지표에서 Qwen3-VL-8B 19.7%, InternVL3.5-8B 13.3% vs 인간 98.2%). TOMATO, MMIU 등 선행 벤치마크도 프레임별 인식과 시간 통합의 괴리를 일관되게 보고하며, 이는 본 과제에 과제 특화 학습과 순서 편향을 다루는 추론 설계가 필요함을 시사한다.

2.4 TPRU

TPRU [11](2026)는 텍스트 조건부 프레임 재정렬(3–4프레임, 순열 출력)로 본 과제와 거의 동일한 설정에서 Qwen2.5-VL-7B를 SFT하여 MuirBench Ordering에서 GPT-4o를 추월했다(34.4% vs 23.4%). 따라서 이 사실에 근거하여 Qwen3-VL-8B를 SFT하는 방향성을 결정할 수 있었다.

3 Method

3.1 EDA

EDA 과정에서 가장 먼저 눈에 들어왔던 것은 No ordering 컬럼이었다. 이 컬럼이 있다는 것은 분명 학습에 이용할 수 있는 정보라고 생각해 확인해본 결과, 학습 데이터의 약 15.5%는 정답이 $[1, 2, 3, 4]$인 비셔플 표본임을 확인할 수 있었다.

따라서 모델은 가능한 24가지 순열 중에서, $[1, 2, 3, 4]$라는 정답을 15.5% 가량 가져가는 prior를 가져가야 한다고 생각하였다. 이는 추후에 항등 헤드 보정을 떠올리는 데에 도움이 되었다.

또한 EDA 중 또 눈에 띄었던 정보는, 데이터 중에서 이미지 4장의 소스가 다른 것처럼 느껴지는 데이터들이 존재했다는 것이다. 분명 같은 이미지 4장이 배열되어 있는데, 특정 데이터는 이미지 4장이 같은 영상으로부터 샘플링된 이미지처럼 보였고, 특정 데이터는 그렇지 않은 것처럼 보였다. 이에 프레임 해상도와 이미지 형식을 분석한 결과, 일부 데이터들은 다른 영상으로부터 추출한 프레임을 섞어서 사용했다는 것을 알 수 있었다. 실제로 모델은 균일 해상도 데이터를 더 맞추기 어려워하는 모습을 보였기 때문에, 균일 해상도 데이터와 혼합 해상도 데이터를 나눠서 스코어링하였다. 테스트 데이터에는 학습 데이터에 없던 고해상도 AI 생성형 이미지도 포함되어 있어 로컬 검증 결과를 유형별로 해석하였다.

3.2 베이스 모델과 순열 후보 스코어링

먼저 베이스 모델을 선정하기 위해 두 후보의 Zero-shot 성능을 체크했다. 동일한 검증 표본에서 Qwen3-VL-8B-Instruct의 exact match는 0.371로 InternVL3.5-8B의 0.283보다 8.8%p 높았고 속도도 약 14배 빨랐다. RTX 3090에 배포 가능한 32B 4-bit 모델도 시험했으나 8B bf16보다 정확도가 3.4%p 낮았다. 이에 성능, 속도, 메모리를 종합하여 Qwen3-VL-8B-Instruct를 선택하였다.

각 순열 후보 π의 점수는 캡션과 네 프레임을 prefix로 입력한 뒤 정답 문자열의 토큰 로그확률을 합산하여 계산하였다.

$$S_{\text{forward}}(\pi) = \sum_{t=1}^{T} \log p\left(y_t^{(\pi)} \mid x, y_{<t}^{(\pi)}\right), \tag{1}$$

$$\hat{\pi} = \arg\max_{\pi \in \mathcal{P}_4} S_{\text{forward}}(\pi). \tag{2}$$

이 방식은 항상 가능한 순열 중 하나를 선택하므로 파싱 오류가 없고, 모든 후보의 점수를 후처리와 오류 분석에 활용할 수 있다.

3.3 보조 QA Task and Datasets

전체 Training Data는 76,280개로 구성되며, 메인 순서 과제 47,675개, 보조 QA Task 1 9,535개, 보조 QA Task 2 19,070개로 구성하였다.

먼저 메인 순서 과제의 경우 기본 Training Data를 사용하는 동시에, 동일한 Task를 이미지 Frame이 주어지는 순서만 무작위로 재배치하여 4배 추가로 Augmentation하였다. 이는 모델이 주어진 이미지의 순서와 무관하게 정답을 맞출 수 있도록 하며, 잘못된 자리 편향이 생기는 것을 방지한다. 이때, 증강 후에도 항등 순열 비율을 실제 데이터의 15.5%와 유사하게 유지하여 Prior 정보를 줄 수 있도록 하였다. 보조 QA Task 1의 경우, 원래 Task와는 역방향으로 p(캡션 | 정렬된 프레임)를 학습할 수 있도록, 정렬된 Frame을 주고 캡션을 정답으로 낼 수 있도록 SFT를 진행하였다.

보조 QA Task 1을 채택하게 된 배경은, EMNLP 2019에서 발표된 기계번역의 noisy-channel 재순위 연구 [5]에서 아이디어를 얻었다. 번역 Task에서 후보를 p(번역문 | 원문)이 아니라 p(원문 | 번역문)으로 재검증하듯, 우리 Task에서도 후보 순서대로 프레임을 재배열하고 캡션이 생성될 조건부 확률을 상위 후보(top-4)에 대해 계산하면, 순방향 점수와 오류 구조가 다른 독립 검증 신호로 상호 보완 효과를 낼 수 있을 것이라 생각하였다. 두 점수 모두 동일한 단일 모델에서 계산되므로 별도 모델 출력의 조합(앙상블)에 해당하지 않는다.

실제로 해당 역방향 채점을 학습 전 모델로 시험한 결과 저마진 표본의 top-4 내 정답 선택률은 30%(찬스 25%)에 그쳤지만, 순방향에서 틀리는 문제를 역방향에서는 맞추는 사례가 확인되어(오류 비상관) 상호 보완 효과가 날 수 있을 것이라 판단하였다. 이에 이 능력 자체를 보조 QA Task 1로 학습하였고, 학습 후 선택률은 46%로 상승하였다.

또한 보조 QA Task 2의 경우, Training Data로부터 두 프레임의 선후관계를 묻는 과제로 국소적인 Temporal Causality를 학습할 수 있도록 하였다.

3.4 추론 파이프라인과 최적화

24개 후보를 독립적으로 계산하면 이미지 인코딩과 prefix 계산이 24회 반복된다. 따라서 이미지와 캡션 prefix의 KV 캐시를 한 번 계산한 뒤 모든 후보가 공유하도록 구현하였다. 200개 표본 검증에서 기존 방식과 exact match가 동일했고 표본당 처리 시간은 7.1초에서 2.1초 이하로 감소했으며, 피크 메모리는 약 19.1GB로 RTX 3090 제약에 부합한다.

입력 위치 편향을 줄이기 위해 프레임 슬롯을 순환 이동한 네 개의 TTA 뷰를 사용하였다. 각 프레임이 A, B, C, D 위치에 한 번씩 나타나므로 위치 편향을 평균적으로 상쇄한다.

다음으로 VLM이 생성한 24개 점수 벡터에 세 가지 경량 보정을 가산하는융합 헤드를 적용하였다.

(1) 장면 그룹핑 보너스: CLIP·DINOv2 임베딩과 저수준 통계 피처로 학습한 Logistic Classifier가 프레임 쌍의 "같은 클립" log-odds를 추정하고, 각 후보 순열의 인접 3쌍에 대해 합산한다(λ_u=1.0, 혼합해상도 λ_m=0.5). 같은 클립의 프레임들이 시간상 인접해야 한다는 구조는 혼합해상도 표본 전수에서 성립함을 확인하였다.

(2) 방향 헤드: 프레임 쌍 피처의 반대칭 차이를 입력받는 소형 MLP가 두 프레임의 선후 log-odds를 추정해 인접쌍에 합산한다(λ_d=0.5). 쌍 정확도는 64%로 VLM 자체보다 낮지만 오류가 VLM과 비상관이라 박빙 후보에서 casting vote로 작동하며, 검증 EM을 1.2%p 개선하였다.

(3) 항등 보정: EDA에서 확인한 항등 순열 prior(15.5%)를 반영해 항등 후보에 상수 μ=0.5를 가산한다.

세 보정기의 학습 라벨은 모두 해상도와 정답 순서에서 자동 생성되며, 학습 데이터에 대응 유형이 없는 AI 생성형 시험 표본은 CLIP kNN 거리 기반으로 별도 판별하여 (1)·(2)를 자동 차단한다.

마지막으로 후보 순서대로 프레임을 재배열한 뒤 캡션의 조건부 로그확률을 계산하여 순방향 점수와 결합하였다.

$$S_{\text{final}}(\pi) = S_{\text{forward}}(\pi) + \lambda_r S_{\text{reverse}}(\pi), \qquad \lambda_r = 0.7. \tag{3}$$

역방향 Reranking은 마진 하위 30% 표본의 top-4 후보에만 적용하였다. 커버리지를 100%까지 확대한 검증 실험에서 고마진 표본은 구조적으로 거의 뒤집히지 않아 70% 부근에서 이득이 포화하였으며, 추가 계산을 저확신 표본에 집중하는 것만으로 대부분의 이득을 확보할 수 있었다.

4 Experimental Setup

검증 분할은 학습 데이터에서 층화 추출한 1,000개로 구성하였다. 항등 순열 표본(15.5%)이 전체 정확도를 부풀리므로 전체 exact match와 함께 셔플 표본 exact match를 주지표로 측정하였고, 해상도 층(혼합/균일)별로도 분해하여 해석하였다.

5 Results

5.1 단계별 성능 변화

표 1은 주요 개선 단계의 결과이다. zero-shot 전수 스코어링은 공개 점수 0.710을 기록했으며, LoRA SFT만으로 0.890까지 상승하였다. 이후 TTA, 시각 구조 보정, 다과제 학습, 역방향 Reranking을 결합하여 최종 0.93193에 도달하였다.

표 1: 단계별 성능 변화

방법	검증 전체 EM	검증 셔플 EM	공개 LB
Zero-shot 24순열 전수 스코어링	0.361	0.331	0.710
LoRA SFT (v1)	0.554	0.542	0.890
항등 prior 및 시각 구조 보정	0.598	0.542	0.90401
기본 Task Data Augmentation + TTA + 점수 융합	0.609	0.561	0.910
저마진 선택 연산 + 역방향 Reranking	0.616	0.566	0.912
보조 QA Task 학습 + oversampling 제거	0.618	0.6024	0.926
전체 데이터 full-fit (최종)	–	–	**0.93193**

표 2: 주요 ablation과 채택 여부

실험	선택	핵심 결과
TTA 뷰 수 $k = 1, 2, 4$	$k = 4$ 채택	$k = 4$에서 전체 EM 2.6%p 증가
TTA $k = 8$	기각	융합 성능 0.613에서 0.594로 하락
균일층 8배 oversampling	기각	자기-EM 0.955, 검증 일반화 하락
Random/Hard-negative DPO	기각	검증 및 공개 LB에서 일관된 이득 없음
32B 4-bit 모델	기각	8B bf16 대비 전체 3.4%p 하락
캡션 역방향 Reranking	채택	저마진 후보에서 상보적 개선

5.2 Ablation 결과

6 Analysis

24순열 전수 스코어링은 문제의 유한한 출력 구조를 직접 반영한다. 자유 생성에서는 순서를 이해해도 설명 문장이나 잘못된 형식을 출력할 수 있지만, 후보 채점은 항상 가능한 순열 하나를 반환한다. 또한 최종 검증 오답의 90.3%가 top-1과 top-2 마진 2 미만 구간에 집중되었고 마진 4 이상인 확신 오답은 2.2%뿐이었다. 이는 모델이 어려운 표본을 비교적 잘 식별하며 저마진 표본에만 추가 계산을 투입하는 전략이 합리적임을 보여준다.

TTA를 순환 4뷰로 구성한 것은 이론과 실증 양쪽의 근거를 따른 결정이다. 순환 이동 4개는 각 프레임이 네 입력 슬롯을 정확히 한 번씩 방문하는 최소 완전 집합으로, 위치 편향을 구조적으로 상쇄하는 가장 저렴한 구성이다. 실제로 뷰 수 스윕에서 k:1→2→4로 늘릴수록 수익이 증가했으나(전체 EM $+0.9 \rightarrow +2.6$%p), 역순환 4뷰를 추가한 k=8은 새로운 편향을 상쇄하지 못한 채 결합 점수만 희석시켜 오히려 하락했다(0.613→0.594). 즉 k=4는 "더 많은 뷰"가 아니라 "편향 상쇄에 필요충분한 뷰"라는 기준으로 선택된 값이다.

융합 헤드가 유효한 이유도 정확도가 아니라 독립성에 있다. 방향 헤드의 쌍 정확도는 64%로 VLM 자체의 방향 판단(약 93%)보다 훨씬 낮음에도 검증 EM을 1.2%p 올렸는데, 이는 두 신호의 오류가 비상관이어서 VLM이 확신하지 못하는 박빙 후보에서 casting vote로 작동하기 때문이다. 같은 이유로 보정 가중 λ는 작게 유지되어 고마진 표본을 뒤집지 않고 박빙에서만 개입한다. 흥미롭게도 모델이 강해질수록 외부 보정의 최적 가중은 줄어들었는데(zero-shot에서 최적 λ=6 → SFT 후 0.5∼1.0), 이는 보정 헤드의 역할이 "모델이 못 하는 것을 대신하는 것"에서 "박빙의 균형을 깨는 것"으로 옮겨감을 보여준다.

역방향 스코어링은 "캡션과 프레임으로 순서를 선택하는" 순방향 판단을 "해당 배열이 캡션을 얼마나 잘 설명하는가"라는 다른 조건부 확률로 검증한다. 두 점수는 같은 모델에서 계산되지만 오류 구조가 완전히 같지 않아 박빙 후보에서 상보적으로 작동하였다. 캡션 생성 보조 과제 적용 후 역방향의 top-4 내 정답 선택률이 약 30%에서 46%로 높아진 결과는 학습 과제와 추론 전략의 연결이 효과적이었음을 보여준다.

데이터 설계에서는 단순 oversampling보다 일반화 진단이 중요했다. 가장 어려운 균일해상도 층을 8배 반복했을 때 rank 32 모델의 학습 표본 자기-EM은 0.955까지 올랐지만 검증 성능은 하락하였다. 반복 노출이 일반화가 아닌 암기를 일으킨다고 판단하여 oversampling을 제거했고, 균일층 정확도는 0.468에서 0.505로 회복되었으며 공개 점수도 0.912에서 0.926으로 상승하였다. 이 진단은 어댑터 용량 선택과도 상호작용했다: oversampling이 있던 조건에서는 rank 32가 rank 16보다 열세로 보였으나(암기가

용량에 비례해 증폭되므로), 데이터를 고친 뒤에는 rank 32의 표현력이 온전히 발현되어 최종 채택되었다.

층별 성능은 장면 경계가 뚜렷한 혼합해상도 표본 약 0.92, 균일해상도 표본 약 0.505로 차이가 컸다. 남은 병목은 서로 다른 장면의 구분보다 하나의 연속 동작 안에서 손의 위치, 물체 이동량, 자세 변화와 같은 미세 시간 단서를 비교하는 문제이다. 향후에는 모든 프레임 쌍의 상대 순서를 일관되게 추정하거나 변화량 중심의 시각 표현을 추가할 필요가 있다.또한 학습 데이터에 대응 유형이 없는 AI 생성형 시험 표본은 로컬 검증이 제한되었다는 한계가 있다.

7 Conclusion

본 연구는 캡션 기반 네 프레임 순서 복원 문제를 24클래스 순열 스코어링으로 재정의하고, Qwen3-VL-8B LoRA 학습에 셔플 증강, 방향 쌍 QA, 캡션 생성 과제를 결합하였다. 추론에서는 KV 캐시로 반복 계산을 제거하고 순환형 TTA와 역방향 캡션 Reranking을 적용하였다. 그 결과 공개 리더보드 점수를 0.710에서 0.93193으로 향상시켰고, 피크 메모리 19.1GB와 약 2.7시간의 추론 시간으로 RTX 3090 제한을 충족하였다. 모델 크기를 단순히 늘리거나 어려운 표본을 반복하는 것보다 출력 구조를 직접 활용하고 저마진 구간에 선택적으로 계산을 투입하는 전략이 효과적이었다. 향후에는 균일한 단일 장면의 미세 시간 순서와 학습 데이터에 없는 AI 생성형 이미지에 대한 별도 검증 체계를 강화해야 한다.

참고 문헌

[1] Qwen Team, "Qwen3-VL-8B-Instruct Model Card," 2025.

[2] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR, 2022.

[3] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.

[4] M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," TMLR, 2024.

[5] K. Yee et al., "Simple and Effective Noisy Channel Modeling for Neural Machine Translation," EMNLP-IJCNLP, 2019.

[6] I. Misra et al., "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification," ECCV, 2016.

[7] H.-Y. Lee et al., "Unsupervised Representation Learning by Sorting Sequences," ICCV, 2017.

[8] D. Xu et al., "Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction," CVPR, 2019.

[9] Y. Song et al., "Burn After Reading: Do Multimodal Large Language Models Truly Capture Order of Events in Image Sequences?," Findings of ACL, 2025.

[10] B. Li et al., "TimeBlind: A Spatio-Temporal Compositionality Benchmark for Video LLMs," arXiv:2602.00288, 2026.

[11] Z. Gao et al., "TPRU: Advancing Temporal and Procedural Understanding in Large Multimodal Models," ICLR, 2026.