SNU AI Challenge 참가 결과 보고서

Team: 고구마숙주돈가스

한양대학교 정보공학전공 김동희

한양대학교 데이터사이언스전공 임지훈

1. 개요

본 과제의 목표는 주어진 스토리라인(캡션)에 맞게 4개의 이미지 프레임을 함께 분석하여 프레임의 올바른 순서를 추론하는 것이다.

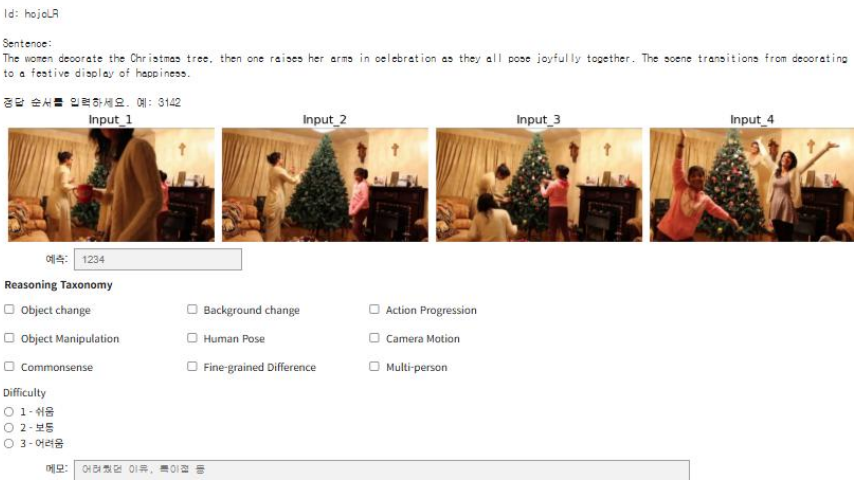
본 연구의 핵심 기여는 세 가지이다. 첫째, 네 프레임의 전역적 일관성을 직접 학습하는 Direct Ordering을 중심 과제로 설정하여 pair 예측을 단순 결합할 때 발생하는 순환·모순을 줄였다. 둘째, Screen-48과 Clean-152의 2단계 검증으로 checkpoint 선택의 계산량과 신뢰도를 함께 관리하고, 선택된 Direct 모델에 Pairwise 예시를 보조적으로 연속학습하여 전체 순서와 미세 선후 관계를 동시에 보완하였다. 셋째, 테스트 전체를 반복 추론하지 않고 사전 검토된 소수 사례에만 실행 기록을 추적할 수 있는 strict TTA 규칙을 적용해 추론 비용과 예측 변경 범위를 제한하였다.

2. 초기 데이터 분석

데이터의 정답 순열 분포를 확인한 결과, [1,2,3,4]가 1,478개로 가장 많았고 나머지 23개 순열은 각각 305~386개로 대체로 유사하지만 완전히 균등하지는 않았다. 또한 [1,2,3,4]인 1,478개는 모두 No_ordering=True인 사례였다. 이에 따라 No_ordering 사례가 특정 정답에 집중된 데이터 불균형 가능성을 확인하였다.

초기 데이터 분석에서는 데이터를 언어 정보인 캡션과 시각 정보인 이미지 프레임으로 나누어 접근하였다.

우선 human baseline을 만들어 사람이 캡션과 클립을 관찰하며 순서를 결정할 때 사용하는 시각적, 언어적 단서를 기록하였다.



캡션 문장 분석에서는 시간 표현을 나타내는 특정 단어들을 통해 문장의 경계를 만들어 사건을 분리할 수 있다. 이 과정을 통해 각 이미지 프레임이 캡션의 어떤 내용과 대응되는지 비교하고 모델이 이미지 프레임의 시간 흐름을 더욱 용이하게 판단할 수 있도록 한다.

이미지 프레임 분석에서는 사람과 객체의 위치 변화, 물체의 등장과 소멸, 카메라 이동 등 사람이 순서를 판단할 때 활용하는 시각적 단서를 human baseline을 통해 유형화하였다. 또한 실패 가능성이 높은 사례를 분류하여 어떤 점을 중점적으로 모델을 제작해야 할지 정리하였다.

[표1] 시간 순서 추론에 활용한 주요 시각 단서와 예상 난이도

유형	주요 시간 단서	예상 난이도
객체 이동	물체나 사람의 위치 및 이동 궤적 변화	중간
미세 동작	손, 시선, 자세와 같은 국소적 변화	높음
카메라 이동	배경 전체의 평행 이동 또는 회전	높음
장면 전환	화면 전체에서 발생하는 급격한 변화	중간
등장·사라짐	객체가 화면 안팎으로 이동하는 변화	중간
반복 행동	유사한 동작이 여러 차례 반복되는 현상	매우 높음
텍스쳐 부족	특징점과 시각적 단서가 부족한 장면	매우 높음

3. 모델 설계 및 개선 과정

초기 베이스라인은 별도의 설정 없이 Qwen2.5-VL-7B-Instruct 모델을 기반으로 하여 4개의 이미지와 캡션을 입력하여 전체 시간 순서를 한 번에 추론하였다.

본 연구에서는 모델의 학습 방식을 Direct Ordering과 Pairwise Ordering 두 가지 방안으로 구분하였다.

Direct Ordering은 전체 이미지 프레임을 동시에 입력받아 [1,2,3,4]와 같은 전체 순열을 출력하고, Pairwise Ordering은 2개의 이미지 프레임을 한 쌍으로 하여 둘 중 시간 순서가 우선인 이미지를 예측하는 방식이다.

우선 Qwen7B을 NF4 4-bit로 양자화하고 LoRA를 적용한 후 이미지 프레임과 캡션을 이용하여 전체 순열을 출력하도록 모델을 학습시켰다. 이 단계에서는 Direct Ordering으로만 학습을 진행하였으며, 학습하지 않은 초기 베이스라인 모델보다 큰 폭으로 성능이 향상되었다. 그 후 프레임의 제시 순서를 바꾸어 추가 학습을 진행해 보았지만 영향이 미미했다.

별개로 Pairwise Ordering만으로 모델을 학습시켜보았다. 6쌍의 pair 예측을 조합하여 전체 순서를 출력하였는데 이 시도에서는 두 프레임 사이의 선후 관계는 잘 학습하였으나 전체 순서 예측에는 어려움을 겪었다. 따라서 Pairwise Ordering 방식을 단독으로 사용하기보다 Direct Ordering을 보조하는 형태를 이루어야 한다는 방향성을 설정하였다. 그 후 검증 데이터를 제외한 학습 데이터에서 Direct 학습 예시를 늘려가며 추가 학습을 진행하였다.

추가적으로 모델이 어려워하는 사례를 중심으로 학습하도록 프롬프트를 조정하였는데 실제 예측이 변경된 데이터가 적어 개선 효과가 미미하였다.

Qwen 기반 학습 외에도 모델의 성능을 보완하기 위해 다양한 접근을 추가로 실험하였다.

프레임 간 시각적 연속성을 분석하기 위한 방법을 탐구하는 과정에서「Depth from Defocus 데이터셋 수집을 위한 이미지 정합 알고리즘 개발」[1]의 이미지 정합 접근을 참고하였다. 이를 본 과제에 맞게 적용하는 과정에서

SIFT와 BFMatcher를 이용해 프레임 사이의 특징점을 대응시키고 RANSAC으로 잘못된 매칭을 제거하여 인접한 프레임을 탐색하고자 하였다.

먼저 모든 프레임 쌍에 대해 SSIM 기반 사전 유사도와 SIFT·RANSAC 특징을 계산하였다. 이후 사전 유사도, RANSAC score와 CLIP 이미지 유사도를 조합해 쌍별 우선순위를 산출하고, 계산 비용이 큰 optical flow는 우선순위가 높은 3쌍에만 적용하였다. 평가는 실제 정답에서 인접한 프레임 쌍의 탐색 능력, 두 프레임 사이의 시간 거리와 RANSAC score 사이의 관계성, 그리고 전체 프레임의 순서 탐색 능력을 고려하였다. 이때 정방향과 역방향을 구별할 수 없기에 무방향 연결 순서로 평가하였다.

하지만 SIFT와 RANSAC 기반의 점수만으로 전체 프레임의 순서를 예측했을 때의 효과는 미미하였다. 따라서 이를 독립적인 모델로 사용하는 대신 예측의 보조 특징으로 활용하는 방향을 설계하였다. 최종 제출 모델에는 포함되지 않았지만 실패 가능성이 높은 사례를 분석하는 과정에 이를 참고하였다.

다음으로는 이미지와 캡션의 대응 관계를 활용하기 위해 CLIP 모델 기반의 예측을 실험하였다. 사전학습된 CLIP 모델의 가중치를 고정하고 이미지 프레임과 캡션 특징을 결합하는 모듈을 학습하여, 6개 프레임 쌍 각각의 선후 관계를 이진 예측하도록 하였다. 추론 시에는 이 6개 pair의 선후 확률을 이용해 가능한 24개 순열을 점수화하였다. 또한 Qwen의 인접 순열에만 CLIP을 적용하는 방법도 고려해보았다. 하지만 CLIP 기반 방법의 성능이 미미하여 주된 모델의 보조 자료로만 사용하였다.

Ablation 결과, V1 Direct LoRA는 초기 베이스라인보다 Validation Exact가 0.1450에서 0.3700으로 상승했고 Public Score도 0.39965에서 0.69458로 개선되어 Direct 학습의 효과가 가장 분명했다. 반면 V2의 순열 continuation은 Validation을 0.0100 높였지만 Public과 Private 점수가 모두 하락하여 표시 순서 증강이 곧바로 일반화 성능으로 이어지지는 않았다. V5와 R2에서는 어려운 사례만 반복하기보다 일반 사례를 함께 균형화했을 때 Public Score가 최대 0.72425까지 상승했다. CLIP 결합은 이 수준을 넘지 못했으므로 최종 설계는 여러 모델의 결과를 결합하기보다 Qwen 기반 Direct 표현을 안정적으로 학습하는 방향을 선택하였다.

[표2] 모델 구성별 Validation 및 Leaderboard 성능 비교

실험 구성	주요 변경 사항	Validation Exact	Public	Private
초기 Qwen7B Baseline	학습 없이 전체 순열 추론	0.1450	0.39965	0.36178
V1 Direct LoRA	Direct Ordering 학습	0.3700	0.69458	0.69512
V2 Permutation Continuation	프레임 표시 순서 변경 추가 학습	0.3800	0.67713	0.67886
V4 All-train Continuation	전체 train pool 추가 학습	미측정	0.69982	0.69918
V5 Hard Mining	어려운 사례와 규칙 프롬프트 적용	0.3650	0.71902	0.71951
R2 Hard-Balanced	어려운 사례와 일반 사례의 균형 조정	미측정	0.72425	0.71951
CLIP Pairwise Fusion	CLIP pair 예측과 Qwen 결과 결합	별도 평가	0.71029	0.71138
CLIP Fusion	CLIP 신뢰도 기반 Qwen 결과 결합	별도 평가	0.72076	0.71544
Direct 1,000·step250	Direct 학습 데이터 확대	별도 평가	0.70680	0.69105

4. 최종 모델 및 제출 파이프라인

최종 모델은 Qwen2.5-VL-7B-Instruct를 기반으로 구성하였다. 네 장의 프레임 전체 순서를 직접 예측하는 Direct 학습을 중심으로 사용하고, 두 프레임의 선후 관계를 판단하는 Pair 학습을 보조적으로 적용하였다. 제한된 GPU 환경에서도 모델을 학습할 수 있도록 모델 가중치를 NF4 4-bit 형식으로 압축하고, 전체 모델 대신 작은 LoRA adapter만 학습하는 QLoRA 방식을 사용하였다. 또한 학습 안정성과 메모리 효율을 고려하여 bfloat16 연산과 SDPA Attention을 적용하였다. LoRA 설정은 rank 16, alpha 32, dropout 0.05로 통일하였다.

전체 데이터 중 200개는 학습에 사용하지 않고 시드 고정 후 검증용으로 따로 분리하였다. 이때 검증 데이터 200개 중 48개는 여러 학습 단계의 모델을 빠르게 비교하는 Screen Validation으로 사용하였다. 나머지 152개는 최종 후보 모델의 성능을 더 정확하게 비교하는 Clean Validation으로 사용하였다.

나머지 학습 데이터에서는 네 장의 프레임 전체 순서를 맞히는 Direct Ordering 학습 예시 1,000개를 구성하였다. 모델은 총 250번의 optimizer update를 수행하였으며, 25 step마다 중간 모델인 checkpoint를 저장하였다. 먼저 Screen-48에서 성능이 높은 checkpoint 3개를 고른 뒤, 이 모델들을 Clean-152에서 다시 평가하였다. 그 결과 step175 모델이 예측 성능이 가장 좋아 최종 Direct 모델로 선정되었다.

이후 step175 모델을 바탕으로 추가 학습을 수행하였다. PairAux 추가학습 예시 200개 중 네 장의 전체 순서를 맞히는 Direct 예시를 160개(80%), 두 장의 선후 관계를 판단하는 Pairwise 예시를 40개(20%)로 구성하여 총 50 step을 더 학습하였다. Direct 예시의 비중을 높게 유지해 기존의 전체 순서 예측 능력이 떨어지지 않도록 하고, Pairwise 예시를 통해 서로 비슷한 프레임의 세부적인 앞뒤 관계를 더 잘 구분하도록 하였다.

테스트 단계에서는 Pair Auxiliary 모델을 사용해 전체 819개 샘플의 기본 예측을 생성하였다. 이후 사전 검토된 strict-39 계획을 적용하였다. 39개 대상 중 22개는 기본 예측이 이미 사전 정의한 대안 순열과 같아 유지하였고, 3개는 기록된 기준 예측과 일치하지 않아 추가 추론을 생략하였다. 나머지 14개에 대해서만 프레임 표시 순서를 네 가지 방식으로 바꾸어 추가 추론하였으며, 사전 정의한 대안 순열이 네 번 중 세 번 이상 선택된 2개 샘플의 예측을 교체하였다. 이를 통해 기본 예측을 최대한 유지하면서 검토된 일부 샘플만 제한적으로 수정하였다.

최종 제출 결과 Public Score는 0.72251, Private Score는 0.69512를 기록하였다.

[표3] 최종 PairAux 및 strict-39 TTA 제출 성능

실험 구성	주요 변경 사항	Validation Exact	Public	Private
최종 PairAux	Direct 중심 + Pairwise 보조학습	**0.4408**	**0.72425**	0.69512
PairAux + strict-39 TTA	일부 샘플에 4-view 후처리	**해당 없음**	0.72251	**0.69512**

※ strict-39 TTA는 테스트 데이터 전용 후처리이므로 Validation Exact를 별도로 산출하지 않았다.

5. 성능 향상 및 오류 분석

선택된 step175 Direct 모델의 Clean-152 Exact는 0.4211이었고, PairAux 연속학습 후 0.4408로 0.0197 상승하였다. Pairwise accuracy도 0.76425에서 0.78070으로 개선되어 Pairwise 예시가 Direct Ordering의 보조 과제로 유효함을 확인하였다. 반면 strict-39 TTA는 Public Score가 0.72425에서 0.72251로 소폭 하락하고 Private Score는 0.69512로 동일했다. 따라서 이 후처리는 점수 향상 기법이라기보다 불확실한 예측의 변경 범위를 제한하는 보수적 규칙으로 해석해야 하며, 향후에는 별도 검증셋에서 대상 선택 기준과 투표 임계값을 먼저 검증할 필요가 있다.

오류는 주로 반복 행동, 손·시선 같은 미세 동작, 텍스쳐가 부족한 장면에 집중되었다. 이 경우 인접 프레임의 시각적 차이가 작아 Pairwise 관계가 불안정하고, SIFT·RANSAC은 프레임 간 연결성은 측정할 수 있어도 시간 방향을 판별하지 못했다. CLIP 역시 정적인 이미지-문장 대응에는 유용하지만 미세한 시간 진행을 표현하는 데 한계가 있었다. 제안 방법의 장점은 전역 순서를 직접 최적화하면서 제한된 GPU에서도 재현 가능한 학습·선택 절차를 제공한다는 점이며, 한계는 200개 검증셋의 크기와 strict 대상 선정에 수작업 검토가 포함된다는 점이다.

6. 결론

본 연구에서는 무작위로 배열된 네 장의 이미지 프레임을 올바른 순서로 추론하기 위해 Qwen2.5-VL-7B-Instruct 기반의 학습, 추론 모델을 구축하였다. 초기 모델의 성능을 개선하기 위해 여러 방법을 고안하였고 그 중 네 프레임의 전체 순서를 직접 예측하는 Direct Ordering이 가장 안정적이었다. 또한 Pairwise Ordering은 Direct 학습을 보완하는 보조 자료로 활용하였을 때 효과적이었다.

최종 모델은 Direct Ordering 학습 후 Pairwise 예시를 추가로 학습한 구조로 구성하였다.

하지만 프레임 사이의 변화가 작거나 반복 행동이 포함된 경우에는 이미지와 캡션만으로 정확한 선후 관계를 판단하기 까다로웠다. 또한 SIFT·RANSAC과 CLIP 기반 방법에서 한계가 확인되어, 추후에는 보다 안정적인 추론을 가능하게 하는 새로운 방법론을 탐구할 필요가 있다.

참고 문헌: [1] 안성민, 최동일, “Depth from Defocus 데이터셋 수집을 위한 이미지 정합 알고리즘 개발,” 대한기계학회논문집 A, 49(1), 11–16, 2025. doi:10.3795/KSME-A.2025.49.1.011