문장-이미지 시간 정렬과 순열 TTA를 활용한 이미지 순서 복원

팀명: HYperBall

팀원: 임재민 최민석 최준혁

1. 문제 정의 및 접근 개요

본 과제의 목표는 하나의 문장과 순서가 섞인 이미지 네 장이 주어졌을 때, 문장이 설명하는 시간 흐름에 맞게 이미지의 순서를 복원하는 것이다. 이미지가 네 장이므로 가능한 정답은 총 24개이다. 겉으로는 24개 순열 중 하나를 고르는 분류 문제로 볼 수 있지만, 실제로 정답을 찾으려면 (1) 문장의 시간 구조 이해, (2) 각 이미지의 시각적 상태 파악, (3) 문장과 이미지 상태의 정렬, (4) 인접 프레임 간 선후관계 검증이 동시에 필요하다. 예를 들어 문장이 컵을 들어 올린 뒤 책상에 내려놓는 과정을 설명한다면, 컵을 들고 있는 장면과 컵을 옮기는 장면의 선후관계를 구분해야 한다.

초기 데이터 분석과 실험을 통해 두 가지 문제를 확인하였다. 첫째, 각 이미지를 독립적으로 캡셔닝한 뒤 설명을 정렬하는 방식은 이미지 사이의 미세한 상태 차이를 놓치기 쉬웠다. 둘째, 모델이 실제 시간 관계보다 처음 제시된 이미지 순서인 [1, 2, 3, 4]를 그대로 답하는 경향이 있었다. 이는 모델이 이미지 내용뿐 아니라 입력 위치와 프레임 라벨에도 영향을 받을 수 있음을 보여준다.

이를 해결하기 위해 최종 방법을 sentence-chain 기반 QLoRA 학습, multiscale Label TTA, identity 보호 선택적 추론의 세 단계로 구성하였다.

2. 모델 구조와 핵심 아이디어

2.1 백본 모델과 QLoRA

본 과제는 문장의 시간 구조와 각 이미지의 시각적 상태를 함께 이해하고 정렬해야 하므로 생성 모델 기반으로 접근하였다. 백본으로 멀티모달 생성 모델 Qwen3.6-27B를 사용하였다. 32GB GPU에서 학습하기 위해 백본을 4-bit NF4와 double quantization으로 로드하고, BF16으로 연산하는 QLoRA를 적용하였다. LoRA는 언어 모델의 attention 및 MLP projection 계층에 적용했으며, rank 8, alpha 16, dropout 0.05로 설정하였다. 학습 파라미터는 전체의 0.1454%로, 사전학습 표현을 보존하면서 과제에 필요한 부분만 조정하였다.

2.2 Sentence-chain 프롬프트

sentence-chain 프롬프트의 핵심은 문장을 이미지에 대한 부가 설명이 아니라 시간 흐름을 정의하는 조건으로 취급하는 것이다. 이를 위해 문장이 기술하는 사건의 순서를 명시적으로 따르도록 다음 지침을 포함하였다. 첫째, then, before, after, finally 등의 표현은 실제로 보이는 사건을 연결할 때에만 선후 관계 제약으로 사용하고, while, as와 같이 동시성을 나타낼 수 있는 표현은 무조건적인 순서 제약으로 해석하지 않도록 하였다. 둘째, 문장에 네 개보다 적은 사건만 명시된 경우에는 준비, 시작, 진행, 결과, 접촉, 물체 이동, 자세 변화와 같은 미세한 시각적 진행을 근거로 같은 사건 내부의 프레임을 정렬하도록 하였다. 셋째, 입력으로 제시된 순서가 이미 정답일 가능성을 인정하되 그 이유만으로 우대하거나 배척하지 않도록 하여 [1,2,3,4] identity 편향을 완화하였다. 마지막으로 최종 순열을 확정하기 전에 인접한 프레임 쌍의 선후 관계를 다시 검증하도록 하여 인접 프레임 간 오정렬을 줄이고자 하였다.

2.3 핵심 아이디어

① 네 이미지를 각각 독립적으로 설명한 뒤 정렬하는 대신, 문장의 사건 흐름과 네 이미지의 시각적 상태를 직접 대응

② 학습에 앞서 데이터를 정제하고 품질에 따라 손실 가중치를 차등 적용하여 명백한 잡음과 학습이 어려운 표본을 다르게 처리

③ 동일한 이미지를 여러 순서와 라벨로 제시하고, 예측을 원본 이미지 번호로 복원하여 투표하는 Label TTA로 입력 위치 및 라벨 편향을 완화

④ 낮은 해상도와 높은 해상도의 상호보완적인 예측을 결합하고, 안정적인 identity 예측은 보호하면서 나머지 표본에만 reversed 추론을 추가

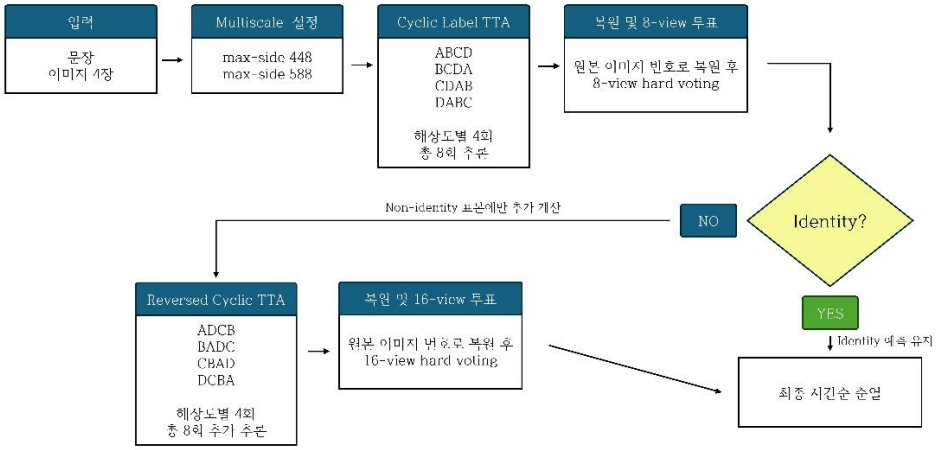

그림 1. Identity-protected multiscale Label TTA 추론 파이프라인. 두 해상도의 cyclic 예측을 결합하고, identity가 아닌 표본에만 reversed TTA를 추가한다.

3. 학습 과정

3.1 데이터 정제 및 품질 기반 손실 가중치 적용

학습 데이터에서 강한 잡음은 제거하되, 단순히 어려워 보인다는 이유만으로 많은 데이터를 버리지 않는 것을 원칙으로 하였다. 전체 9,535개 표본을 이미지 파일 해시, 문장 시작 형태, 프레임 유사도 기준으로 점검하였다. 먼저 한 표본 안에 파일 해시가 동일한 프레임이 포함된 33개를 제거하였다. 동일한 프레임이 두 번 이상 등장할 경우 모델의 의미 있는 순서 정보 학습에 방해가 될 수 있기 때문이다. 문장이 then, and then, after that, next, finally 등으로 시작하는 181개는 앞 문맥이 잘린 문장일 가능성이 있다고 보았다. 일부 정보는 여전히 학습에 활용될 수 있기에 해당 샘플들은 제거하지 않고 손실 가중치를 0.25로 설정하였다. 또한 네 프레임을 32×32 grayscale로 축소한 후 모든 프레임 쌍의 평균 절대 오차(MAE)를 계산하였다. 가장 다른 프레임 쌍조차 MAE가 0.075 이하인 119개는 이미지 구분이 매우 어려운 표본으로 보고 손실 가중치를 0.5로 설정하였다. 마지막으로 전체 평균 손실의 크기가 과도하게 줄어들지 않도록 손실 가중치 평균을 1로 정규화하였다. 정규화를 통해 데이터 품질에 따른 중요도는 유지하면서 전체 학습 규모는 기존과 동일하게 유지했다.

3.2 학습 설정

하이퍼파라미터	값
batch size	1
gradient accumulation	8
optimizer	AdamW
learning rate	2e-5
scheduler	cosine
warmup ratio	0.03
gradient clipping	0.3

No_ordering 여부에 따라 비율이 유지되도록 층화하여 전체의 10%인 954개를 holdout validation으로 분리하였다. 나머지 데이터에 앞서 설명한 품질 검사를 적용한 뒤 1 epoch 학습하였다. 이미지 입력의 총 픽셀 수는 최소 $256 \times 28 \times 28$, 최대 $448 \times 28 \times 28$ pixels 범위로 제한했다. 더 높은 설정에서는 일부 배치에서 Out Of Memory(OOM) 오류가 발생했기 때문에 시각 정보량과 학습 안정성을 고려하여 해당 범위를 최종 설정으로 채택하였다.

4. 추론 과정과 성능 향상 기법

4.1 Cyclic Label TTA

모델이 시각적 시간관계를 충분히 확인하지 않고 첫 번째 위치나 A 라벨을 시작 장면으로 선택할 가능성이 있다. 이를 줄이기 위해 ABCD, BCDA, CDAB, DABC의 네 cyclic view를 사용하였다. 같은 원본 이미지가 네 번의 추론에서 A, B, C, D 라벨을 각각 한 번씩 갖게 된다. 네 view의 결과를 원본 번호로 복원하고 가장 많이 나온 순열을 최종 답으로 선택한다. 예를 들어 서로 다른 view에서 [B, C, A, D]와 [A, B, D, C]가 출력되더라도 원본 번호로 복원하면 모두 [2, 3, 1, 4]가 된다. 이 때 모델이 이미지 내용에 근거해 일관되게 판단할수록 동일한 원본 순열에 표가 모이게 된다.

4.2 Multiscale 결합

각 cyclic view를 max-side 448과 588에서 각각 추론하여 샘플당 총 8개의 표를 얻었다. 기존 평가 이미지 대부분은 원본 긴 변이 448보다 컸기 때문에 두 설정이 다른 크기의 입력을 만들었다. 448은 이미지를 더 축소하여 장면 전체의 큰 변화에 집중하게 하고, 588은 작은 객체나 미세한 자세 변화를 더 많이 유지한다. 따라서 두 해상도를 결합했을 때 서로의 오답 일부를 보완할 수 있다.

재현 코드에서는 데이터의 원본 해상도가 달라져도 같은 원칙으로 max-side를 선택하기 위해 다음 규칙을 사용한다. High side는 학습 당시 상한에 가까운 588과 전체 이미지 긴 변 중앙값 중 작은 값으로 정한다. Low side는 기존에 검증한 비율 448/588을 유지한다. High side에서 중앙값을 사용한 이유는 일부 극단적으로 큰 이미지가 평균을 지나치게 높이는 것을 막기 위해서이다.

$$High\ side = \min\left(588, median_i(\max(W_i, H_i))\right)$$

$$Low\ side = round(High\ side \times \frac{448}{588})$$

4.3 Reversed TTA 와 identity 보호

Cyclic view는 시작 위치는 바꾸지만 이미지가 배열된 진행 방향은 유지한다. 모델이 왼쪽에서 오른쪽으로 제시된 배열 자체에 영향을 받을 수 있으므로, ADCB, BADC, CBAD, DCBA의 reversed cyclic view도 추가하였다. 이 경우에도 각 출력은 원래 이미지 번호로 복원한 뒤 투표한다.

다만 validation을 분석했을 때 cyclic multiscale이 identity를 예측한 표본의 정확도는 80.92%로 비교적 높았다. 모든 표본에 reversed 결과를 더하면 이미 안정적으로 맞힌 identity가 다른 순열로 바뀌는 경우가 있었다. 따라서 먼저 두 해상도의 cyclic 결과 8개로 baseline을 만들었다. Baseline이 identity이면 그대로 최종 답으로 사용하고, identity가 아닌 경우에만 두 해상도의 reversed 결과 8개를 추가해 총 16표로 다시 결정하였다.

5. 실험 결과 및 분석

5.1 프롬프트 비교 실험

동일한 954개 holdout에서 학습 프롬프트를 비교하였다. Legacy는 일반적인 순서 정렬 지시문, task-aware는 문장이 목표 순서를 정의한다는 점을 강조하였고, Sentence-chain은 여기에 사건과 상태 매칭과 인접 프레임 상호관계 검토를 추가하였다. 추가로 Sentence-chain v2는 같은 사건 내부의 미세한 진행 상태와 인접 프레임 전이 검토를 보강하고, 입력된 순서가 정답인 경우도 다른 순열과 동일한 기준으로 평가하도록 하였다.

학습 프롬프트	정답 수	Exact accuracy	Position accuracy
Legacy	560 / 954	58.70%	71.54%
Task-aware	561 / 954	58.81%	71.51%
Sentence-chain	564 / 954	59.12%	71.54%
Sentence-chain v2	**566 / 954**	**59.33%**	**71.60%**

프롬프트에 과제의 목적을 구체적으로 반영할수록 greedy 정확도가 점진적으로 향상되었다. 특히 Sentence-chain v2는 가장 높은 성능을 보여 최종 어댑터에 사용하였다.

5.2 해상도 비교 실험

두 단일 scale을 결합한 cyclic multiscale이 단일 추론보다 높은 성능을 기록하였다. 전체 표본에 reversed TTA를 적용하면 성능이 소폭 하락했지만, 정확도가 높은 identity 예측을 보호하는 선택적 결합으로 최고 정확도 61.11%를 얻었다. 따라서 본 실험에서는 단순히 추론 횟수를 늘리는 것보다, 신뢰도가 높은 예측을 보호하면서 추가 추론을 선택적으로 적용하는 것이 더 효과적이라는 결론을 내릴 수 있었다.

추론 방법	예측 수	정답 수	Exact accuracy	Position accuracy
448 cyclic	4-view	568 / 954	59.54%	71.51%
588 cyclic	4-view	560 / 954	58.70%	71.07%
448+588 cyclic multiscale	8-view	577 / 954	60.48%	72.30%
Cyclic + reversed multiscale	16-view	576 / 954	60.38%	72.25%
Identity-protected hybrid	**선택적 8/16-view**	**583/954**	**61.11%**	**72.80%**

5.3 오류 분석

최종 모델은 holdout 954개 중 583개를 정확히 맞혀 61.11%의 exact-match accuracy를 기록하였다. 오답 371개 중 138개는 단 하나의 이미지 쌍 관계만 잘못 판단한 근접 오류였다. 따라서 모델이 시간적 흐름 전체를 이해하지 못했다기보다, 서로 상태가 비슷한 두 프레임의 인접 순서를 반대로 결정하면서 exact match를 놓치는 경우가 주요 오류 유형임을 알 수 있다.

문장 길이에 따른 정확도 차이도 컸다. 10단어 이하 표본의 정확도는 39.25%, 11~20단어는 33.33%였지만, 21~30단어는 69.68%, 31~40단어는 83.99%였다. 또한 문장에 명시적인 시간 단서가 있는 표본은 70.77%, 없는 표본은 37.86%였다. 문장이 짧거나 시간 표현이 없으면 모델이 이미지 변화만으로 순서를 추론해야 하므로 난도가 높아졌다.

시각적 유사도가 가장 높은 25% 표본의 정확도는 44.35%였지만, 프레임 차이가 가장 뚜렷한 25%는 73.64%였다. 프레임 사이의 차이가 작을수록 정확도가 크게 낮아졌고, 특히 물체의 위치, 개수, 개폐 상태처럼 작은 변화만 존재하는 샘플에서는 인접한 두 프레임의 순서를 뒤집는 오류가 자주 발생하였다.

6. 장점과 한계

제안 방법의 가장 큰 장점은 문장과 이미지를 별도로 처리하지 않고, 문장이 설명하는 시간 흐름에 이미지 상태를 직접 대응시킨다는 점이다. Label TTA는 추가 학습 없이도 입력 위치 및 라벨 편향을 줄일 수 있었고, Multiscale은 서로 다른 수준의 시각 정보를 결합하며, identity 보호 방식은 불필요한 추가 추론과 안정적인 정답 변경을 줄일 수 있었다. 데이터 정제에서도 강한 노이즈는 확실히 제거하고, 어려운 표본은 제거하지 않고 품질에 따라 영향력을 조절하였다.

반면 계산 비용은 큰 한계이다. Non-identity 표본은 최대 16회의 모델 생성이 필요하므로 단일 greedy 추론보다 시간이 오래 걸린다. 기법에서의 한계도 존재한다. Identity 보호는 validation에서 관찰한 특성을 이용했기 때문에 분포가 크게 달라진 데이터에서 낮은 정확도를 보일 수 있다. 또한 원본 이미지의 해상도가 낮거나 인접 프레임이 거의 같으면 높은 해상도로 확대하더라도 새로운 정보가 생기지 않는 점도 존재한다.

7. 결론

본 과제에서는 이미지 순서 복원을 문장이 설명하는 사건 흐름과 이미지의 시각적 상태를 정렬하는 문제로 접근하였다. Qwen3.6-27B에 QLoRA SFT를 적용하고, sentence-chain 프롬프트를 통해 사건과 상태를 비교하는 기준을 학습하였다. 추론에서는 동일한 이미지를 여러 순서와 해상도로 반복해서 보고 원본 번호 기준으로 투표하는 Label TTA를 사용하였다. 여기에 reversed TTA와 identity 보호를 결합하여 정확도와 계산 효율을 함께 고려하였다. 그 결과 Public 점수를 greedy 기준 0.883에서 최종 0.917까지 향상시켰다. 이후 개선에서는 짧은 문장과 미세한 인접 상태 변화의 선후관계를 정확하게 판단하는 방법을 찾는 것이 중요하다고 결론지었다.