1. 문제 정의 및 접근 개요

본 과제는 주어진 문장 설명과 순서가 섞인 4장의 프레임 이미지를 바탕으로, 원래 영상 내 사건의 시간적 순서를 복원하는 것을 목표로 한다. 각 샘플은 하나의 문장과 Input_1부터 Input_4까지의 네 개 이미지로 구성되며, 모델은 네 이미지가 실제로 어떤 순서로 발생했는지 예측해야 한다.

이 과정에서 모델은 사람의 자세 변화, 물체의 이동, 손과 물체의 접촉 여부, 행동의 시작과 종료 상태 등 여러 프레임에 걸쳐 나타나는 변화의 흐름을 종합적으로 이해해야 한다. 또한 문장에 제시된 사건 설명을 각 프레임의 시각적 단서와 연결해야 하므로, 이미지와 텍스트를 함께 해석하는 멀티모달 추론 능력이 요구된다.

본 연구에서는 해결해야 할 난점을 두 가지로 정리하였다. 첫째, 네 프레임의 순서를 결정하는 것은 개별 이미지를 독립적으로 인식하는 문제가 아니라, 프레임 사이의 관계와 전체 사건의 흐름을 동시에 판단하는 문제이다. 따라서 이미지를 각각 인코딩하거나 순서 관계를 부분적으로 나누어 예측하는 방식은 전체 맥락을 충분히 반영하지 못해 일관된 순서를 도출하기 어렵다. 둘째, 문장이 시간 순서에 관한 단서를 충분히 제공하는 샘플과 단일 연속 동작을 담아 프레임 간 미세한 시각적 차이에 의존해야 하는 샘플이 함께 존재한다. 특히 후자의 경우 프레임들이 시각적으로 유사해 순서 판단의 불확실성이 커질 수 있다.

이에 네 프레임을 하나의 입력으로 함께 제시하고, 전체 사건의 순서를 직접 생성하도록 모델을 학습하였다. 또한 초기 추론 결과가 일관되지 않은 어려운 샘플을 선별하여 추가적으로 판단하는 추론 전략을 설계하였다. 본 연구에서는 생성된 각 순서의 시퀀스 로그확률을 개별 예측 수준의 신뢰도로 활용하고, 서로 다른 입력 순열에서 얻어진 예측들의 일치도를 샘플 수준의 신뢰도로 활용하였다. 전자는 투표가 동률인 경우 후보를 구분하는 데 사용하였으며, 후자는 추가 TTA24 수행 여부를 결정하는 데 사용하였다. 구체적인 모델 구조와 학습 및 추론 방법은 다음 절에서 설명한다.

2. 모델 구조 및 학습 방법

2.1 Base Model: Qwen3-VL-8B

본 실험에서는 base 모델로 Qwen3-VL-8B-Instruct를 사용하였다. Frame Ordering Task는 캡션으로 주어진 문장 정보와 네 장의 shuffled frame 이미지를 함께 해석해야 하므로, 텍스트와 이미지를 동시에 입력받을 수 있는 대형 Vision-Language Model이 필요하다. 이에 본 팀은 이미지-텍스트 통합 추론이 가능한 Qwen3-VL 계열 모델을 선택하였고, 기존 4B 모델 대비 더 큰 표현력을 가진 8B 모델을 사용하였다.

2.2 QLoRA 기반 Fine-tuning

8B 규모의 Vision-Language Model을 전체 파라미터 기준으로 파인튜닝하는 것은 Colab 환경에서 GPU 메모리 부담이 컸다. 이를 해결하기 위하여 본 실험에서는 QLoRA 기반 파인튜닝을 적용하였다. QLoRA는 base model을 4-bit로 양자화한 상태에서, 일부 모듈에 LoRA adapter를 추가하여 학습하는 방식이다. 이를 통해 전체 모델을 직접 학습하는 것보다 훨씬 적은 메모리로 task-specific 학습을 수행할 수 있었다.

2.3 Direct Output 형식 설계

본 실험에서는 모델이 긴 reasoning 과정을 생성하도록 하는 CoT 방식 대신, 최종 순서만 직접 출력하는 Direct Output 방식을 채택하였다. 이는 생성 길이를 줄이고, 출력 parsing의 안정성을 높이기 위함이다. Direct Output 방식은 모델이 중간 설명이나 추론 과정을 길게 출력하지 않고, 최종 프레임 순서만 바로 출력하도록 설계하는 방식이다. 모델은 캡션과 4개의 프레임을 입력 받은 뒤, 가장 이른 프

레임부터 가장 늦은 프레임까지의 순서를 JSON 형식으로 출력하도록 학습된다.

시스템 프롬프트는 모델에게 shuffled frame 4장을 보고 실제 순서를 판단한 뒤, {"order":[f1,f2,f3,f4]} 형태로만 답하도록 지시한다. 학습 target 역시 JSON order만 포함하도록 구성되어 있다. 위 방식은 불필요한 설명 생성을 줄이고, 최종 submission으로 변환하기 쉬운 구조를 만든다는 장점이 있다. 또한 parsing 실패 가능성을 줄여 추론 단계에서 더 안정적으로 다수결 또는 log-probability 기반 voting을 적용할 수 있다.

2.4 전체 학습 데이터 9,535개 활용

기존 실험에서 Qwen3-VL-8B와 Direct Output, QLoRA 조합의 성능이 충분히 확인되었다고 판단하여, 본 실험에서는 별도의 validation split을 두지 않고 전체 학습 데이터를 모두 사용하여 학습 데이터 활용량을 극대화하는 전략을 채택하였다. 이러한 결정은 선행 실험들(4B, 8B 계열)에서 로컬 검증 점수(0.51~0.54)가 Public 점수(0.88~0.90)보다 항상 낮았으나 상대적 변화는 대체로 일치했던 경향에 근거한다. 검증셋은 절대 성능보다 실험 간 상대 비교에만 유효했으므로, 모델 구성이 이미 검증된 시점에서는 전량 데이터를 학습에 투입하는 것이 합리적이라 판단하였다.

또한 AUG=2 설정을 통해 각 샘플의 프레임 순서를 무작위로 섞는 shuffle augmentation을 적용하였다. 이를 통해 모델이 입력 프레임의 고정된 위치에 의존하지 않고, 다양한 frame order 상황에서도 올바른 시간 순서를 예측하도록 학습되었다. 또한 Colab 환경의 메모리 제약을 고려하여, 실제 학습에서의 물리적 batch size는 1이지만, gradient accumulation을 통해 유효 batch size 8에 해당하는 학습 효과를 얻도록 구성하였다.

2.5 이미지 입력 해상도 예산 확대: max pixels 384 → 512

Frame Ordering Task에서는 프레임 간 차이가 매우 미세한 경우가 많다. 물체와의 접촉 여부, 인물의 자세 변화처럼 작은 시각적 단서가 시간 순서 판단에 중요한 역할을 할 수 있다. 따라서 이미지 입력 해상도 예산을 높이면 모델이 더 많은 시각 정보를 활용할 수 있다. 본 실험에서는 기존 설정 대비 이미지 입력 예산을 확대하여 MAX_PIXELS를 384 기준에서 512 기준으로 증가시켰다.

3. 추론 전략

3.1 Cyclic TTA4 구성

학습 과정에서 shuffle augmentation을 적용하였음에도 추론 시 모델이 입력 프레임의 물리적 위치에 따라 다른 판단을 내릴 가능성을 완전히 배제할 수 없다. 이를 보정하기 위해 자기일관성(self-consistency) 기반 TTA를 적용하되, 무작위 순열 대신 [0,1,2,3], [1,2,3,0], [2,3,0,1], [3,0,1,2]의 순환(cyclic) 순열 4가지를 사용하였다. 이를 통해 4개 프레임 각각이 입력 위치 1~4에 정확히 한 번씩 배치되도록 보장하여, 특정 위치에 대한 모델의 편향이 추론 결과에 미치는 영향을 최소화하였다. 무작위 샘플링과 달리 매 추론마다 동일한 4가지 관점을 일관되게 확보할 수 있다는 장점이 있다.

3.2 다수결 voting의 한계

3.1에서 얻은 4회의 추론 결과를 단순 다수결로 집계할 경우, 표가 균등하게 갈려 우세한 답이 나오지 않는 상황에서는 임의의 기준으로 하나를 선택할 수밖에 없다. 이러한 동률 상황은 모델이 실제로 확신하지 못하는 애매한 샘플에서 빈번하게 발생하며, 단순 다수결은 이 경우 모델이 가진 생성 확신도를 전혀 활용하지 못한다는 한계가 있다. 특히 4회라는 적은 투표 수에서는 이러한 상황이 구조적으로 발생하기 쉬워, 3.5~3.8에서 다룰 표본 재분류 및 추가 투표 전략으로 이어진다.

3.3 개별 예측 수준 신뢰도: 다수결 기반 집계 및 Sequence Log-probability Tie-break

이러한 한계를 보완하기 위해, 다수결에서 동률이 발생한 경우에 한해 각 후보 답의 시퀀스 로그확률합, 즉 생성 확신도를 비교하여 최종 답을 결정하였다. 여기서 시퀀스 로그확률은 하나의 입력 순열에서 생성된 개별 예측이 모델 내부에서 얼마나 높은 확률을 부여받았는지를 나타내므로, 개별 예측 수준의 신뢰도로 사용하였다. 이는 모델이 특정 순서를 생성할 때 부여하는 확률이 높을수록 해당 예측의 신뢰도가 높다는 전제에 기반한다. 이때 동일한 답을 지지하는 여러 표의 확신도를 합산함으로써, 단일 추론의 확신도보다 여러 순열에 걸쳐 일관되게 지지받는 답을 우선하도록 하였다. 다수결이 명확한 경우에는 기존 결과를 유지하고, 동률 상황에서만 추가 정보를 활용하는 보수적인 설계이다.

3.4 출력 parsing 실패 대응

Greedy decoding 결과가 유효한 JSON 형식으로 파싱되지 않는 경우, temperature 0.7, top_p 0.9 샘플링으로 1회 재시도하였다. 이때 재시도 시드를 샘플 ID와 순열 조합의 해시로 고정하여, 샘플링 재시도임에도 실행 간 결과가 완전히 재현되도록 하였다. 파싱 실패로 표가 줄어들면 다수결과 확신도 기반 tie-break 모두의 신뢰도가 저하되므로, 이러한 재시도는 3.1~3.3에서 구성한 투표 체계 전체의 안정성을 지탱하는 역할을 한다. 재시도 이후에도 파싱에 실패하는 경우에는 해당 표를 제외하고 남은 표로 판정하며, 모든 표가 무효화된 극단적인 경우에 한해 기본 순서를 반환하도록 하여 예외 상황에서도 항상 유효한 출력을 보장하였다.

3.5 샘플 수준 신뢰도: TTA4 Vote Consistency 기반 분류

3.1~3.4에서 얻은 TTA4 결과에 대해, 서로 다른 입력 순열에서 동일한 예측이 얼마나 일관되게 나타나는지를 샘플 수준의 신뢰도로 정의하였다. 이는 3.3에서 사용한 개별 생성 시퀀스의 로그확률과는 다른 개념으로, 각 예측 자체의 확률이 아니라 네 번의 TTA 결과가 서로 얼마나 일치하는지를 측정한다. TTA4 결과에서 네 순열 모두 유효한 표를 얻었고(파싱 성공), 그중 최다 득표가 3표 이상인 경우(즉 4:0 또는 3:1)에만 모델이 뚜렷한 확신을 보인 쉬운 샘플로 간주하였다. 표가 최다 2표 이하로 분산되거나(2:2, 2:1:1, 1:1:1:1), 파싱 실패로 유효표가 4개 미만인 경우는 판정이 불확실한 애매한 샘플로 분류하였다. 파싱 실패 자체를 불확실성 신호로 취급하여 보수적으로 확장하는 설계이다. 이는 3.3에서 활용한 개별 시퀀스의 생성 확신도와는 별개로, 여러 순열에 걸친 예측의 일관성 자체를 표본 수준의 확신도로 활용한 것이다. 이 기준은 3절에서 이미 계산된 투표 결과만으로 판단 가능하여 별도의 추가 연산이 필요 없다는 점에서 효율적이다.

3.6 쉬운 샘플과 애매한 샘플 분리

이 기준에 따라 전체 테스트 샘플을 쉬운 샘플(4:0, 3:1)과 애매한 샘플(2:2 이하 또는 유효표 부족)로 분리하였다. 쉬운 샘플은 이미 충분한 확신을 얻었으므로 추가 연산 없이 TTA4의 결과를 그대로 채택하고, 애매한 소수의 샘플에만 추가 연산을 투입함으로써 전체 추론 비용을 크게 늘리지 않으면서도 개선 여지가 큰 부분에 자원을 집중할 수 있었다. 이러한 선별적 접근은 전체 샘플에 대해 24개 순열을 모두 추론하는 방식과 비교할 때, 계산 비용을 크게 절감하면서도 애매한 샘플에 한해서는 동일한 수준의 세밀한 판정이 가능하다는 이점이 있다.

3.7 애매한 샘플에 대한 추가 20개 permutation 추론

애매한 샘플에 대해서는 3절에서 계산된 4가지 순환 순열의 결과를 그대로 재사용하고, 여기에 나머지 20가지 순열을 추가로 추론하여 총 24가지 전체 순열에 대한 결과를 확보하였다. 순열 수를 늘릴수록 특정 답에 대한 표가 우연이 아닌 일관된 근거를 바탕으로 수렴할 가능성이 높아지므로, 애매한 샘플에서 더 정확한 판정을 내릴 여지가 커진다. 이를 통해 해당 샘플에 한해서는 가능한 모든 프레임 배치 조합을 반영한 판정을 내릴 수 있었다.

3.8 다수결 우선 + log-probability tie-break

24회의 추론 결과에 대해서는 다수결을 우선 적용하고 3.3과 동일하게 다수결이 동률인 경우에 한해 생성 확신도를 기준으로 최종 답을 결정하였다. 표본이 4개에서 24개로 늘어나면서 특정 답으로 표가 수렴할 가능성이 높아지므로, 로그확률 기반 tie-break이 실제로 개입하는 빈도는 3절 단계보다 낮아진다. 이처럼 쉬운 샘플과 애매한 샘플에 서로 다른 연산량을 배분하는 전략을 통해 하방 위험 없이 상방 개선만을 노리는 방식으로 설계하였다.

4. 실험 결과 및 분석

4.1 주요 실험 방법 비교

방법	핵심 변경	Public Accuracy
SigLIP permutation scorer	임베딩 기반 24순열 점수화	0.31239
Qwen2.5-VL-3B + QLoRA	생성형 VLM 파인튜닝 도입	0.76090
Qwen3-VL-8B + CoT	모델 규모 확대 및 추론 과정 생성	0.89877
Qwen3-VL-8B + Direct + cyclic TTA4	전체 데이터, 512 해상도, 직접 출력	0.90401
+ Log-probability tie-break 및 재시도	추론 신뢰도와 출력 안정성 활용	0.91273
+ Adaptive TTA24	불확실한 샘플만 24개 순열로 확장	0.91448

초기 임베딩 기반 접근은 프레임 간 전체 사건 흐름을 충분히 반영하지 못해 제한적인 성능을 보였으며, 생성형 VLM을 파인튜닝하면서 성능이 크게 향상되었다. 이후 8B 모델에 CoT 방식을 적용한 결과는 0.899 수준에서 정체되었으나, Direct Output을 적용해 0.90 이상의 성능을 달성하였다. 최종 성능 향상은 모델 구조를 추가로 복잡하게 변경하기보다, 로그확률과 예측 일관성을 활용하여 불확실한 샘플에 추론 연산을 선택적으로 배분하는 과정에서 얻어졌다.

4.2 Public/Private 리더보드 결과

구분	Public	Private
점수	0.91448	0.89837

Public 대비 Private 정확도는 0.01611 하락했으나, 리더보드 내 상대 순위는 두 계단 상승하였다. 이는 평가 데이터가 달라진 상황에서도 제안한 방법이 경쟁력을 비교적 안정적으로 유지했음을 보여준다.

4.3 Ablation 분석

CoT 학습 과정에서 loss가 학습 초반 약 90 step 이내에 0.03 수준까지 급격히 감소하는 현상이 관찰되었다. 이는 정답 JSON 형식 자체의 낮은 entropy에서 기인할 수도 있으나, 모델이 정답을 암기했을 가능성도 배제할 수 없었다. 이를 검증하기 위해 CoT 계열 내에서 서술 다양화와 정답 토큰 가중치 조정을 시도하였으나, Public 점수는 0.899에서 0.902로 거의 변화가 없었다. 이는 관찰된 빠른 loss 감소가 암기보다는 출력 형식의 낮은 entropy에서 기인함을 시사하며, 학습 방식의 미세 조정만으로는 CoT의 성능 정체를 해소하기 어렵다는 판단으로 이어졌다.

4.4 Error Analysis

본 실험에서는 별도의 validation split 없이 전체 학습 데이터를 사용했기 때문에, test set의 정답 기반 오답 분석에는 한계가 있다. 따라서 3.5에서 정의한 TTA4 결과의 vote consistency를 모델 확신도의 간접 지표로 활용하여 오류 발생 가능성이 높은 샘플의 특성을 논의한다. 표가 분산되어 불확실 샘플로 분류되는 경우는 대체로 5.1에서 다루는 시각적 모호성에서 기인하는 것으로 관찰되었다. Adaptive TTA24는 이러한 샘플에만 추가 연산을 배분함으로써 오류 가능성이 높은 영역에 선택적으로 대응하는 역할을 하였다.

5. 한계 및 결론

5.1 한계점

본 실험은 public leaderboard에서 상당히 높은 성능을 보였으나 몇 가지 한계가 존재한다. 첫째, Adaptive TTA24는 모든 샘플에 동일하게 적용되는 방식보다 효율적이지만, 애매한 샘플에 대해서는 추가 추론을 수행하므로 전체 추론 시간이 증가한다. 특히 ambiguous sample의 비율이 높아질 경우 추론 비용이 예상보다 커질 수 있다. 따라서 성능 향상 폭과 계산 비용 사이의 균형을 고려해야 한다.

둘째, TTA 결과의 투표 분포가 항상 실제 정답 여부를 완벽하게 반영하는 것은 아니다. 예를 들어 4:0 또는 3:1로 표가 모였더라도 모델이 공통적으로 잘못된 방향으로 확신할 수 있다. 반대로 2:2처럼 표가 갈리는 샘플이 반드시 어려운 샘플이라고 단정할 수도 없다. 따라서 Adaptive TTA는 실용적인 confidence 추정 방식이지만, 완전한 불확실성 측정 방법은 아니라는 한계가 있다.

셋째, Frame Ordering Task 자체의 어려움도 남아 있다. 프레임 간 차이가 매우 미세하거나, caption이 시간적 단서를 충분히 제공하지 않는 경우에는 대형 VLM이라도 전후 관계를 혼동할 수 있다. 특히 같은 배경에서 인물의 자세만 조금씩 바뀌는 경우, 손과 물체의 접촉 여부가 흐릿한 경우, 동작의 시작과 끝이 명확하지 않은 경우에는 모델이 잘못된 순서를 선택할 가능성이 있다.

5.2 결론 및 향후 개선 방향

본 연구에서는 문장 설명과 순서가 섞인 네 프레임으로부터 사건의 시간 순서를 복원하기 위해 Qwen3-VL-8B 기반 생성형 Vision-Language Model을 활용하였다. QLoRA를 통해 제한된 GPU 환경에서도 8B 모델을 효율적으로 파인튜닝하였으며, 프레임 순서를 무작위로 변형하는 shuffle augmentation과 Direct JSON 출력 형식을 적용하여 입력 위치 편향과 출력 불안정성을 줄였다. 추론 단계에서는 cyclic TTA4를 통해 각 프레임이 모든 입력 위치에 균등하게 배치되도록 하였고, 다수결 동률 상황에서는 sequence log-probability를 활용하여 후보를 구분하였다. 또한 TTA4의 vote consistency를 샘플 수준의 불확실성 지표로 사용하여, 예측이 분산된 일부 샘플에만 나머지 입력 순열을 추가하는 Adaptive TTA24를 제안하였다. 해당 방식은 전체 테스트 샘플 중 7.94%에만 추가 추론을 수행하고 최종 예측의 2.81%만 변경하면서, Public Score를 0.91273에서 0.91448로 향상시켰다.

실험 결과, 프레임을 독립적으로 인코딩하거나 pairwise 관계로 분해하는 접근보다, 네 프레임과 캡션을 공동으로 처리하여 전체 순서를 직접 생성하는 방식이 효과적이었다. 또한 최종 성능 구간에서는 모델 구조를 크게 변경하는 것보다, 기존 강한 생성 파이프라인을 유지하면서 불확실한 샘플에 계산을 선택적으로 배분하는 추론 전략이 유효하였다.

향후에는 독립 validation set을 유지한 통제 실험을 통해 각 구성 요소의 기여도를 정량화하고, vote consistency 외에 entropy, normalized sequence likelihood 또는 calibrated confidence를 결합하여 더 정확한 불확실성 추정기를 설계할 수 있다. 또한 미세한 접촉 변화와 자세 변화를 더 잘 포착할 수 있는 video-native representation이나 temporal encoder를 VLM과 결합하는 방향도 고려할 수 있다.