텍스트로 풀어보는 장면의 재구성

You reached your current quota

서주원(Team Leader), 정우준, 한성재, 박성현

1. 동기와 핵심 기여 (Motivation & Contributions)

본 과제는 자연어 caption 과 시간 순서가 shuffle 되어 있는 4 개의 사진에 대해 원래 시간순을 복원하는 문제이며, 평가 metric 이 Exact Match 이므로 모델이 전체적인 물리/시간적 흐름을 완전히 이해하는 것이 필수적이다. **따라서 우리는 여러 domain 에 대한 coverage 가 있는 10B 이상 모델을 탐색**하였다. 대회 규정상 open-weighted Model 만 사용 가능하므로 Qwen 이나 Gemma 와 같은 모델 family 를 고려하였고 최종적으로 2026 년 5 월 이전에 나온, 여러 benckmark 에서 SOTA performance 를 낸 DeltaNet 구조 기반 **Qwen3.6-27B 모델을 선택하였다. [1, 2]**

우리는 **외부 Data 사용과 생성형 모델로 학습 Data 의 생성이 제한된 규정 속**에서 무리하게 Train Data 를 생성하여 **Low quality Data 로 인한 artifact 를 학습하거나 domain overfit** 을 마주하기보다, 필요한 증강만 넣고 **Inference harness 에 집중**하여 효율적이고 **robust 한 모델을 만드는 것에 주력**하였다. 실제로 train accuracy 에 기반한 RL DPO finetune, train set 에 기반한 CoT chain 을 통한 multi turn training, No_ordering prior 에 기반한 loss function 설계 등 Train 과정의 시도는 **실패로 종결되었으나 Inference harness 에서 3%p 이상의 정확도** 개선을 얻을 수 있었다. 또한 **우리는 robust 한 config 를 얻기 위해 가장 많은 제출(49 회)와 Local holdout validation 을 수행하였고 의도대로 Public/Private Leaderboard 모두 1 위을 달성할 수 있었다.**

우리가 제시한 아이디어의 **핵심 기여는 다음과 같다.**

• **Test 분포 가정의 최소화:** Unseen data 에 대한 성능을 보존하기 위해 Train/Test set 의 분포 가정(No_Ordering flag)에 맹신하기보다 **임의의 입력이 들어올 수 있다고 가정하고 분포에 대한 가정을 최소화하였다**. 실제로 Train set 대로 15.5% No_Ordering 가정을 주입하였더니 Private Score 에서 -2%p 손실이 발생하였고 위 가정이 맞음을 확인할 수 있다.

• **추가 데이터 생성 비용 0, 대신 full data train:** 데이터 규정을 준수하면서 동시에 고품질 캡션 & 사진 4 개를 만드는 task 는 **어렵다.** 실제로 우리는 두 문제를 and 접속사를 사용해 MixUp 하거나 Multiturn CoT chain 생성와 같은 복잡한 방법부터, RL 을 사용해 다양한 학습을 시도했지만 **SOTA outperform 에 실패하였다**. 따라서 **데이터 생성 비용을 0 으로 줄이고** 적은 시간에 준수한 성능의 모델을 먼저 학습시킨 뒤 **inference harness 를 최적화하여 성능을 개선하는 방법론을 선택하였다.**

• **Train 방법 단순화**: 복잡하고 Overfit-prone 한 방법들보다는 robustness 에 집중하여 **index shuffle 증강, 2726 step 만으로 4bit QLoRA 를 시행**하여 상대적으로 **적은 시간과 VRAM 만 요구**하도록 하였다.

• **Test set 의 난이도에 따른 Inference cost 차등 분배:** 우리 harness 의 **핵심**으로, 단순히 모든 sample 에 동일한 TTA logic 을 할당하기보다 **모델이 어렵게 생각하는** (1 위와 2 위 정답 후보 득표차가 2 표 이하일 때, multi view 를 사용한 심화된 2 단계 TTA 사용) **sample 에 상대적으로 많은 resource 를 동적으로 재분배하였다**. 이 기법은 **실제로** holdout validation set 기준 **Exact Match accuracy (60% -> 63.8%)를 상승시켰다**. 소량의 어려운 sample 에서의 큰 정확도 상승이기 때문에 우리는 **투입한 자원 대비 큰 효과를 거두었다고 결론지었다**.

• **Ablation Study:** 수많은 실험을 통해 통상적인 상황과 다른 제약된 대회 규정, 적은 dataset 과 VRAM 제한 속에서 어떤 방법이 SOTA 가 되는지 **많은 실험과 제출 (실제로 우리 팀의 제출 수가 전체 1 위였다.) 로 정당성을 분석하였다.**

2. 전체 방법론 및 모델 구조 (Method Overview & Architecture)

2.1 Base Model Selection

우리가 선택한 **Qwen3.6-27B 모델은 올해 4 월 21 일에 공개된 Multimodal SOTA 모델이며 Apache 2.0 라이선스의 Open weight 모델이므로** 대회 규정에 부합한다. 우리는 **24GiB VRAM 제한의 inferenc**e 을 충족시키기 위해 nf4 Quantization 을 채택하였으며 성능 보존을 위해 Calibration data selection 에 크게 의존하는 PTQ 보다는 QAT 방법인 **QLoRA** 를 선택하였다. [3, 4] QLoRA 를 통해 nf4 Quantization 과 row-rank adapter ΔW finetuning 을 동시에 수행하되 전체 가중치를 양자화 하는 대신,

Vision Tower 를 제외하여 train data image 에 overfit 되는 것을 억제하고 LM Module 만 양자화 하였다. 추론 때는 미리 패키지로 동봉한 W+ΔW 양자화 가중치를 load 만 한 뒤 바로 inference 를 진행하여 모델 다운로드 시간을 단축하였다.

2.2 Prompt format

시스템 프롬프트를 통해 역할을 부여하고, 캡션과 4 프레임 (최대 512px 이 되도록 resize)을 하나의 유저 프롬프트로 입력시켜 Answer[i] = [a, b, c, d] 형태의 순열을 생성하도록 학습하였고 추론 시에도 마찬가지로 처리하였다.

2.3 Inference TTA

대회 규정상 Ensemble 이 금지되어 있으므로 모델의 robustness 를 보장하기 위해 TTA 를 수행하였다. 먼저 전체 sample 을 batch size = 8 로 **phase 1 24-TTA 에 통과시킨다**. 이때 각 TTA view 의 결과 (득표 분포)를 저장하고 득표차가 일정 이하인, **"모호한 샘플"로 판단되는 경우 phase 2 TTA 를 수행한다**. Phase 2 TTA 의 경우 (대회 규정에 정합하도록) **같은 backbone weight 을 사용하지만 다양한 해상도(448, 576, 600, 704)로 사진을 rescale** 해서 모델이 정밀하게 판단하도록 유도하였다.

3. 학습 과정 (Training)

3.1 데이터 사용

먼저 외부 데이터를 일절 사용하지 **않고 train data 전량 (9535 건)만 사용하여 대회 규정을 준수하였다**. Exploratory Data Analysis**(EDA) 결과 Caption 길이가 다양하고, 극소량의 (33 건) 사진이 중복되는 데이터를 제외하면 육안상 문제는 없었다**. 특이사항으로 **15.5% data 가 No_ordering flag ON & [1, 2, 3, 4] label** 을 가졌다. 이를 **일종의 prior 로 사용하려고 시도**하였지만 모델의 **일반화 성능 저해** (리더보드 점수 하락)를 관측하여 사용하지 않았다. 4! = 24 가지의 순열 경우의 수에 대해 **Index shuffle 을 사용하여 유효 데이터를 24 배로 늘리고 선지의 순서에 대한 불필요한 의존성을 배우지 않도록 하였다.**

3.2 학습 설정

사진의 **side 해상도를 512px 로 제한하였다**. EDA 결과 대부분의 사진들이 **512px, 640px 해상도를 가지는 bimodal distributio**n 을 이루므로 그 중 하나를 최대값으로 선택하는 것이 적절하다고 판단하였다. **512px 를 고른 이유는 다음과 같다.**

- **일반화 성능 평가:** robustness 를 평가하기 위해서 모델이 **'train set 에서 보지 못한 처음 보는 해상도'를 입력으로 받을 때 얼마나 성능이 유지**되는지를 public LB 와 holdout validation 으로 검증하였다. 두 경우 모두 512px 가 640px 를 압도하여 512px 을 최종 선택하였다. 단순히 해상도가 크다고 **무작정 640px 을 선택하지 않은 점**이 우리 팀의 전략이다.
- **Train/Inference VRAM: 해상도가 커지면** activation 점유로 인해 **Inference VRAM 이 증가**하고, 마찬가지로 Train VRAM 이 크게 증가하여 **자원을 많이 사용하게 된다**. 따라서 **비슷한 성능이라면 512px 제한을** 선택하는 것이 더 타당하다.

따라서 512px 제한을 사용해 학습을 진행하였다. **gradient checkpoint** 를 채택해 VRAM 을 크게 낮추어 **A100 기준 10 여 시간, 31.73 GB peak VRAM** 만으로 SFT QLoRA 를 진행할 수 있었다. **또한 30 여 시간의 Single RTX 3090 으로도 동일한 학습을 진행**할 수 있으며 이 버전도 패키지에 포함하였다. 실험을 통해 **1e-4 lr cosine 스케쥴러 & checkpoint 2726** 에서 학습 종료, **gradient accumulation 8, paged AdamW 8bit optimizer** 를 최적 레시피로 찾아서 사용하였다. 최종 제출물은 full train 을 수행하였지만, 후술할 실험 결과 Table 에는 모델 간 성능 비교를 위해 10%의 holdout set 을 제거하여 학습한 후 validation 을 수행하였다. (Ablation study). **앞부분 토큰 (사진 4 장 + 캡션)을 -100 으로 masking 하여 응답한 [a, b, c, d] 부문에 대해서만 token-level cross-entropy(causal LM loss)을 objective function 으로 설정해 학습하였다.**

Epoch	Holdout validation
2	0.621
SOTA 설정과 학습량(∫lr) 동일	0.630
3	0.628
4	0.628

Table 1. 구형 모델(SOTA 이전 버전이므로 val score 가 0.638 이 아님)에 대한 Epoch ablation study

4. 추론 과정과 주요 기법 (Inference Harness)

파이프라인 개요

[Phase 1] 4 프레임을 24 가지(4!) 제시순서로 생성 (24 TTA) → 다수결 -> 동점인 경우 logit 기반 score 로 최초 판정 결정
[접전 검출] 1·2 위 선지의 득표차 ≤ 2 인 표본만 선별 (전체의 ~3%, 23 개)
[Phase 2] 448/576/640/704px 로 24 TTA 절차를 각각 재실행
→ 3/4 이상 다른 선지로 합의 시 판정 교체 후보로 등록 → 512px 결과와 대조해 최종 판정

4.1. Phase 1 – 24 TTA

학습 시 데이터 증강과 유사하게 index shuffle 을 진행하여 4! = 24 가지 view 를 제시하고, 각 뷰의 생성 결과를 **다수결** 투표로 결정한다. 동점일 때는 **logit 을 확률로 변환한 score 로 최초 판정을 정한다**. 이 때 **generate mode** 로 모든 형태의 next token 출력을 허용하여 **빠른 추론**이 가능하도록 하였다. 다행히도 모델이 형식을 잘 학습해서 전체 Test set 추론 과정에서 **한 번도 invalid output 이 발생하지 않았다.**

- **24 TTA 인 이유**: 위치 편향을 피하기 위해서는 모든 순열에 대한 경우의 수를 부여하는 24TTA 가 유일한 해답이다. 그리고, 후술할 **Phase 2 TTA 를 하기 위해서는 24 TTA 결과가 필요하기 때문이다.**
- **실측**: holdout validation 기준 no TTA 0.604 → 4-TTA 0.638 → 24 TTA **0.638(+phase 2 additional accuracy)**. TTA 는 확실히 성능 향상에 크게 기여하였다. 4 TTA 대신 24 TTA 를 사용해서 Phase 2 에 재사용할 수 있도록 하였다.

4.2 접전 영역의 데이터에 대한 분석

Phase 1 은 inference harness 의 훌륭한 baseline 을 담당하지만, **접전 구간(각 후보별 득표차가 적은 구간)의 accuracy 는 약 0.23 으로 전체 accuray 에 비해 크게 낮았다**. 규정 상 **다른 모델 가중치를 활용할 수 없으므로** 우리는 TTA 를 추가하여 모델이 더 다양한 **단서를 관측하도록 Phase 2 단계를 설정하였다**. 모든 Sample 이 아닌 접전 영역의 적은 Sample 만 추가 TTA 를 수행하므로 **2h 정도의 시간만 추가로 소모하여 총 추론 시간 제약(24h) 만족과 accuracy 향상을 동시에 달성할 수 있었다. 이것이 우리 harness 의 독창적인 점이다.**

4.3 Phase 2 – Multi-resolution TTA & Final Verdict

앞서 언급한 것처럼 Phase 1 TTA 는 모든 view 가 같은 512px 렌더링을 공유한다는 단점이 있다. **해상도 resize 로 손실되는 단서** (시계 바늘, 미세한 자세 변화, 해수욕장 지형의 변화)로 인해 모델이 모호한 두 사진 간 순서를 착각하거나, **Vision Token 의 유효 해상도가 달라져서 모델이 특정 크기의 patch** 만 주목하는 문제가 있었다. 따라서 우리 팀은 해상도 축을 조정하여 ,**448/576/640/704px 에서 각각 24-TTA 를 재실행하여 다양한 해상도에 대한 결과를 얻었다.**

마지막으로 우리 harness 의 독창적인 요소는 단순한 다수결이나 prob score (log likelyhood 를 정규화한 확신도) 가 아닌 복합적인 기준으로 overturn 여부를 결정한다는 것이다. 실제로 다음과 같은 기준으로 overturn 여부를 결정한다.

Agree (Score 1 위 == 다수결) / score	판정	직관
Agree, score ratio > 2	**교체**	Phase 2 결과의 명백한 우위
Disagree, score 1 위 > 0.9	기각	Phase 2 결과보다 나은 답 존재
Agree, score ratio < 2	유보	Phase 2 신뢰도가 오차 범위 내
Disagree, score 1 위 < 0.9	**교체**	모델 자체 성능 한계, 다수결인 Phase2 선택

Table 2. Inference harness 의 Phase 2 판정 기준

즉 **Phase 2 TTA 가 Phase 1 결론을 강하게 부정**하거나 **Phase 1 결론 자체의 확신도가 적은** 두 상황에서 교체가 이루어진다.

- Phase 2 TTA 결과가 강한 확신도 (prob score 가 2 위 답안과 2 배 이상)을 가지고, 다수결의 동의를 얻을 때
- Phase 1 결과의 확신도가 적어서 noise 가 적은 Phase 2 결과를 믿는 게 robustness 성능상 유리할 때

후처리의 영향은 holdout validation 기준 **0.638** → 접전 재관측·필터 **0.640** 으로 얼핏 정확도에 큰 영향을 주지 않는 것처럼 보인다. 하지만 후처리가 **극소수의 어려운 sample 에서만 발동되었다**는 점과 후처리의 보수적 설계로 인해 실측 시 **Phase 1 의 정답이 Phase 2 의 오답으로 overturn 되는 경우가 한 건도 발생하지 않았다는** 점에서 Phase 2 후처리는 유의미하다.

로컬 추론 환경 (RTX 3090)에서 Test set 819 개 기준 Phase 1 은 약 7 시간, Phase 2 는 약 2 시간 정도 소요되었다. VRAM 은 22GiB 정도로, **bsz 를 무리하게 키우지 않아** Test set 보다 **큰 해상도의 Data 에 대해서도 OOM 이 발생하지 않도록** 하였다.

5. 실험 결과 및 분석 (Experiments: Road to Final Model, Ablations, Error Analysis)

5.1 Road to Final — 단계별 채택 근거

각각의 SOTA 개발 실험에서 **한 번에 하나의 축만 바꾸고 validation 으로 유효성을 검증하였다.**

단계	변경 사항/비고	Holdout validation	결정
Qwen3-VL-4B, 1ep	베이스라인	0.488	출발점
Qwen3-VL-8B, 3ep	모델 스케일	0.539	채택 (+5.1pp)
+ 균등 순열 증강, 4ep	데이터 증강	0.557	채택
+ 4 - TTA 추론	추론 하네스	0.579	**TTA 축 발견**
Qwen3.5-9B (bf16), 512px	세대 교체	0.602	채택
Qwen3.6-27B (4bit), r16	스케일 ↑ + 양자화	0.630	채택 (+2.7pp)
27B r32 + 24 뷰 TTA	rank·TTA 확장	**0.638**	**최종 학습본**
+ Phase 2 도입	접전 재관측	**0.640**	**최종 시스템**

Table 3. 최종 모델(SOTA) 까지의 주요 개발 성과

주목할 비교는 **9B bf16(0.602) vs 27B 4bit(0.638)**이다. Unsloth 에서는 Qwen3.5 와 Qwen3.6 모델이 큰 4bit 양자화 오차를 유발해 **4bit QLoRA 를 추천하지 않았지만 [5]**, 실측 결과에서는 **4bit QLoRA loss 를 감안하더라도 큰 모델(Qwen3.6)을 고르는 것이 정답이었다.** 우리는 **실측 결과를 중시하여 4bit 27B QLoRA 를 선택하였다**. 이처럼 선행 연구를 맹신하지 말고 **모든 가능성에 대해 최소 한 번이라도 실험을 진행하고 ablation study 위주로 모델을 개발해 나간 것**이 우리의 전략이다.

5.2 Ablations

LoRA rank sweep: r16 이하는 성능에 유의미한 영향(하락)이 발생했으며 r32 이후로는 포화되는 양상을 보였다. 따라서 우리는 rank 를 과도하게 늘리면 adapter parameter 가 증가해 학습 자원이 추가로 소모되므로 **32 를 sweet-spot 으로 선정하였다.**

r16	r32 (채택)	r48	r64
0.630	0.638	0.640	0.638

Table 4. Rank ablation study 와 Holdout validation 점수

Learning rate: 이미 epoc 로 총 학습량을 조절할 수 있으므로, **동일한 총 학습량(∫lr)에서 lr 세기 미치는 영향을 실험하였다.** 실제로 **동일한 ∫lr & lr 을 7e-5 로 줄인 장기 schedule** 에서 **val accuracy 가 0.635 로 하락해서 초기 lr 설정을 채택하였다**.

Epoch (1/2/3/4ep): 0.621/0.638/0.638/0.635 로 holdout validation 기준 3epoch 가 가장 성능이 좋았다. 실제 제출물은 위와 달리 holdout 이 없는 full train 으로 학습하므로 **우리는 총 ∫lr 이 보존되도록 checkpoint 2729 를 최종 제출물로 선택하였다.**

추론 해상도 (512 vs 640): 0.638 vs 0.631 로 앞서 언급한 것처럼 **512px 를 선택한 것이 정당성/성능 측면에서 모두 유리했다.**

5.3 Failed Attempt

우리는 **최종 config 확정 후에도 개선 가능성이 있는 변형을 계속 구현하였으나** 최소가정의 원리에 따라 **유의미한 개선을 이루지 못해 다음 개선안들을 기각**했다. 이 **목록 자체가 "복잡한 기법보다는 정석 + harness"라는 우리의 논리를 뒷받침한다.**

변형	Holdout validation (SOTA: 0.6380)	기각 사유
digit-only loss	0.640	Invalid 출력 생성으로 24TTA 사용 불가
보상 기반 미세조정 (RL)	0.630	Train set reward 에 과적합
dropout 0.1 (기본 0.05)	0.633	과도한 Dropout 으로 성능 저하
응답 내 filler 토큰 학습	0.628	Zero-shot filter noise

캡션 길이 정렬 데이터	0.538	성능 저하

Table 5. 채택하지 않은 기타 개발 시도

Train set 의 noisy 한 성분을 제거한 정제 학습도 시도했다. **자의적인 기준으로 100 여 종의 데이터를 제거한 버전**과, 중복된 사진이 존재하는 **33 여 종의 데이터만 제거하여 동일 레시피로 다시 학습했다**. Validation score 0.636, 0.638 로 filter 전에 비해 (0.638) 이득이 없거나 열세였다. **모든 시도에서 데이터 제거가 domain 감소로 성능 하락에 직결했다**. 이 관찰은 최종 모델을 **train 전량(9,535 건)으로 전부 학습**하는 결정으로 이어진다. 규정상 외부 학습 데이터가 금지된 상태에서 **unseen set 에 대한 일반화 성능을 최대한 보존하고**, **full-data 모델의** 학습 표본 정확도 (이전 953 건에 대한 train accuracy)가 0.795 로 **이미 underfitting 에 가까워 데이터 양을 최대한 늘려야 했기 때문이다.**

5.4 Phase 2 TTA hyperparameter tuning

먼저 같은 config 일 때 Phase 2 의 '접전' 경계 설정에 따른 validation 변화를 측정하였다.

접전	**≤1**	**≤2 (채택)**	**≤3**	**≤4**	**≤6**
정답 개수 변화	+3	**+2**	+2	**−3**	−5

Table 6. 접전 영역의 threshold 득표차에 대한 sweep study 와 Holdout validation 결과

'접전'의 경계를 작은 득표차로 설정하였을 때는 Phase 2 TTA 가 유효하였지만 **큰 득표차 구간까지 침범하는 순간 손실로 반전**된다. **이 경향은** 다양한 해상도 Scale (4 종→6 종)과 Phase 2 합의 기준 (3/4, 4/6, 5/6) 등 hyperparameter 에 대해 **일관되게 나타났다**. 결국 우리는 **접전 구간의 정의만 잘 정하면 다양한 config 에 robust 한 성능을 보장할 수 있다**고 결론지었다. 또한, **추가적으로 흑백/좌우반전/색 변화 등의 TTA 도 실험하였지만 좋은 결과를 얻지 못했다**. (접전 지역 한정 정확도 -4.5%p). 결국 다양한 크기의 patch 를 제공하는 **해상도 변경만 Phase 2 TTA 에 기용하는 것이 올바른 처방이었다.**

6. 장점과 한계 (Strengths & Limitations)

6.1 장점

추론 논리의 일관성: Single backbone 구조로 한 번의 학습과 일관된 기준을 사용하여 전체 test set 을 추론할 수 있다.

적은 학습 비용: 외부 데이터를 사용하지 않아 **데이터 전처리 비용이 0 이고**, full data 에 대한 2.5epoch (2726 checkpoint) 정도의 27B 모델 **QLoRA 1 회만으로 모든 학습 과정을 종결**하였기 때문에 train cost 대비 우수한 성능을 얻을 수 있었다.

24TTA 로 인한 일반화 성능 향상: LLM 특유의 환각과 overfit, 그리고 **순서에 따른 bias 를 피하기 위해 24TTA 는 필수적이다**. 우리 팀은 robustness 를 위해 **추론 시간을 다소 사용하더라도 24TTA 를 선택하여 최종적으로 4%p 정도 정확도를 올렸다**.

6.2 단점

동적으로 바뀌는 추론 비용: Phase 2 TTA 에 통과시키는 "접전 샘플"의 양이 Test set 에 따라 동적으로 결정되므로 총 추론 시간이 일부 달라질 수 있다. **External Data 에서 접전 샘플의 양은 약 10%, Test Data 에서 접전 샘플의 양은 약 3%였지만 만약 더 어려운 set 을 마주한다면 추가 추론 시간이 필요하다.**

24TTA 로 인한 추론 시간 소요: 모델이 generate mode 로도 매 query 에 대해 별도 후처리가 필요 없는 [a, b, c, d]의 결과를 바로 내므로 추론 시간이 빨라 **24TTA 가 대회 규정 추론 시간을 위반하지는 않지만, 9h 정도의 긴 추론 시간이 필요하다.**

7. References

[1] Qwen Team, *Qwen3.6-27B* 모델 카드, Hugging Face, 2026-04-21 공개, Apache-2.0 라이선스.

[2] S. Yang et al., "Gated Delta Networks: Improving Mamba2 with Delta Rule," 2024. (arXiv:2412.06464)

[3] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," *ICLR*, 2022. (arXiv:2106.09685)

[4] T. Dettmers, et al., "QLoRA: Efficient Finetuning of Quantized LLMs," *NeurIPS*, 2023. (arXiv:2305.14314)

[5] Unsloth, "Qwen3.6 - How to Run Locally", https://unsloth.ai/docs/models/qwen3.6