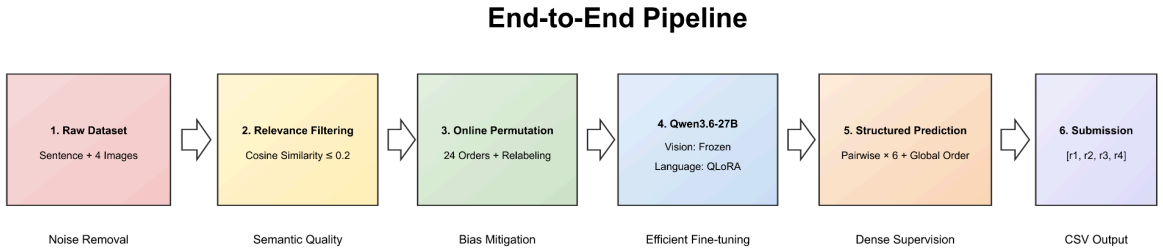
서울대학교 AI Challenge : 텍스트로 풀어보는 장면의 재구성

<Team 도라지>

1. Overall Framework

1-1. Pipeline Overview (Data Preprocessing, Training, and Inference)



단계	내용
① **전처리**	**문장-이미지 코사인 유사도 기반 노이즈 샘플 필터링** (9,535 → 8,744행) **RTX 3090 (24GB) 기준 약 1시간 소요 (API 미사용)**
② **증강**	**온라인 이미지 순열 증강**으로 정답 및 입력 위치 편향 완화
③ **프롬프트**	슬롯 라벨이 부여된 4장 + 문장 + **고정 6쌍 비교 지시문 + 전체 이미지 순서 예측 지시문** 사용
④ **학습**	**Joint Pairwise Temporal Precedence and Global Temporal Order Prediction**으로 학습 신호 강화 QLoRA (r=64) · **RTX 5090 (32GB) 기준 약 12시간 소요**
⑤ **추론**	이미지 크기 리사이징 → Greedy decoding → 정규식 파싱 → 제출 포맷(rank 4개)으로 역변환 **학습 세팅 (4-bit + LoRA)과 동일하게 추론하여 안정성 유지** · **RTX 3090 (24GB) 기준 약 50분 소요**

(cf. 슬롯(slot): 입력 이미지의 위치)

1-2. Model Structure

VLM 백본 모델로는 Qwen3.6-27B 를 선택함:

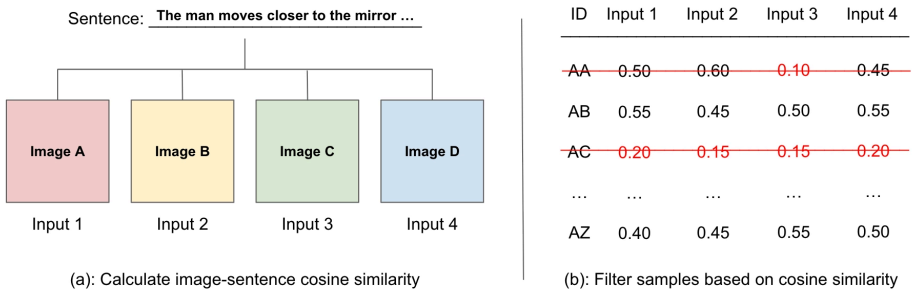
- **Qwen 계열**은 텍스트 모델에 단순히 vision module을 부착한 형태가 아니라, text·image·video token을 interleaved sequence로 구성하여 처음부터 학습한 **네이티브 멀티모달 모델**임.
- 특히 **기존 Qwen3.5 대비 대표적인 멀티모달/비디오 이해 벤치마크에서 전반적인 성능 향상**을 보이므로 본 과제에 적합하다고 판단함 (VideoMME 87.0→87.7, VideoMMMU 82.3→84.4, MLVU 85.9→86.6, MVBench 74.6→75.5)

구성 요소	값
총 파라미터	27.8 B (bf16 기준 55.6 GB)
언어 디코더	64층 · hidden 5120 · GQA (Q heads 24, KV heads 4, head_dim 256)
하이브리드 어텐션	연속된 4층의 어텐션 레이어 중 1층만 표준 attention, 나머지 3층은 게이트 선형 어텐션
MTP 헤드	mtp_num_hidden_layers=1 (multi-token prediction 보조 헤드 내장)
비전 인코더	ViT 계열 depth 27 · hidden 1152 → out_hidden_size 5120 프로젝션

2. Core Idea

2-1. Data Preprocessing: Relevance Filtering

o Motivation: **학습 데이터셋 분석 결과, 화면이 검은색으로 구성되어 시각적 정보가 없는 이미지와 주어진 문장에서 설명하는 장면과 의미적으로 관련성이 낮은 이미지가 일부 포함됨을 확인함.** 이러한 샘플은 이미지의 시간적 순서를 학습하는 데 유효한 정보를 제공하지 못하거나, 잘못된 대응 관계를 학습시키는 노이즈로 작용 가능함.



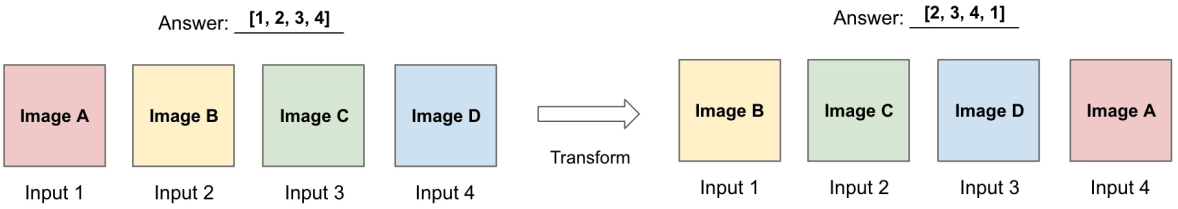
(a): Calculate image-sentence cosine similarity

(b): Filter samples based on cosine similarity

o **Method:** 이미지와 텍스트를 동일한 임베딩 공간에 표현하는 멀티모달 임베딩 모델 Qwen3-VL-8B-Embedding을 사용해 각 샘플에 포함된 네 개의 이미지와 대응 문장을 임베딩한 뒤, 모든 이미지-문장 쌍에 대해 코사인 유사도를 계산함. 이후 하나 이상의 **이미지-문장 쌍에서 코사인 유사도가 0.2 이하로 나타난 샘플을 관련성이 적은 샘플로 판단하여 학습 데이터셋에서 제거함**.

o **Contribution:** 시각적 정보가 부족하거나 이미지-문장 간 의미적 정합성이 낮은 샘플을 제거함으로써 **학습 데이터의 품질을 높이고, 잘못된 이미지-문장 대응 관계로 인해 학습이 방해받는 것을 방지함.**

2-2. Temporal Order Augmentation: Permutation-Based Input Reordering

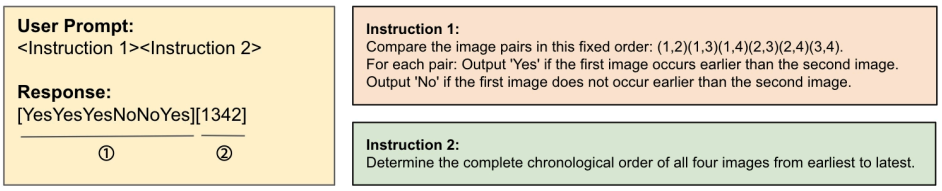


o **Motivation:** 학습 데이터셋의 정답 분포를 분석한 결과, 입력 이미지의 순서가 그대로 시간적순서와 일치하는 [1, 2, 3, 4] 유형의 샘플이 다른 순열에 비해 지나치게 많이 포함되어 있어 **정답 분포의 불균형이 존재함.** 또한, 전체 학습 데이터의 규모가 제한적이기 때문에 모델이 이미지의 시각적 내용과 문장의 의미를 바탕으로 시간 관계를 추론하기보다, **특정 입력 위치가 특정 정답 위치와 자주 대응한다는 위치 편향을 학습할 가능성이 있다고 판단함.**

o **Method:** 학습에서 각 샘플에 대해 **네 개 이미지의 입력 순열을 무작위로 샘플링하고, 선택된 순열에 따라 이미지의 입력 위치를 재배열함.** 이때 재배열된 입력 위치를 기준으로 정답 인덱스를 다시 계산하여 동일한 시간 관계를 보존하도록 정답을 재매핑함.

o **Contribution:** 24개의 가능한 이미지 순열이 보다 균등하게 학습에 활용되어 **정답 순열의 불균형을 완화시킴.** 또한, 모델이 이미지의 절대적인 입력 위치에 의존하는 대신 **이미지 간 시각적 변화와 문장에서 제시된 사건의 전개를 기반으로 시간 순서를 판단하도록 유도함.**

2-3. SFT: Joint Pairwise Temporal Precedence and Global Temporal Order Prediction



SFT : ① Pairwise Temporal Precedence Prediction + ② Global Temporal Order Prediction

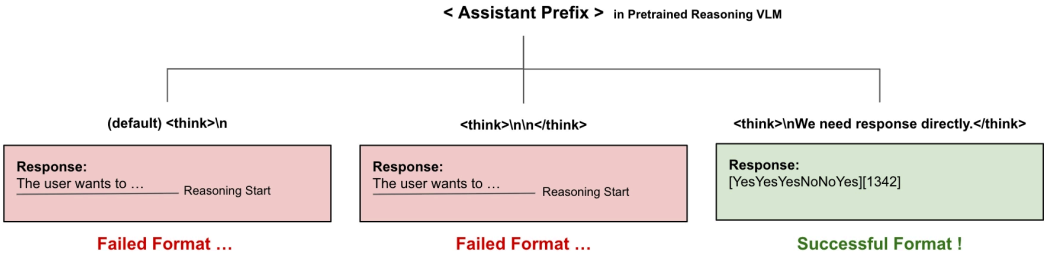
o **Motivation:** Global Temporal Order만을 직접 예측하도록 학습을 진행했을 때, 학습 loss가 일정 수준에서 정체되거나 학습 도중 일시적으로 크게 증가하는 현상이 반복적으로 관찰됨. **전체 정답 순열을 한 번에 생성하는 방식은 최종 순서에 대한 전역적인 지도 신호는 제공하지만, 부분적인 시간 관계 예측 능력을 학습하는 효과는 제한적이라고 판단함.**

o **Method:** 네 이미지로 구성할 수 있는 여섯 개의 이미지 쌍 (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)을 고정된 순서로 제시하고, **각 쌍에 대해 첫 번째 이미지가 두 번째 이미지보다 시간적으로 먼저 나타나는 경우 'Yes' 그렇지 않은 경우 'No'를 출력함.** 이후 동일한 응답에서 네 이미지의 전체 시간 순서를 추가로 출력하도록 학습을 진행함.

o **Contribution:** **전체 시간 순서에 대한 전역적인 지도뿐만 아니라, 추가적으로 두 이미지 사이의 국소적인 선후 관계에 대한 조밀한 학습 신호를 추가함.** 이를 통해 복잡한 전체 순서 예측 문제를 상대적으로 단순한 쌍별 비교 문제로 부분적으로 분해하고,

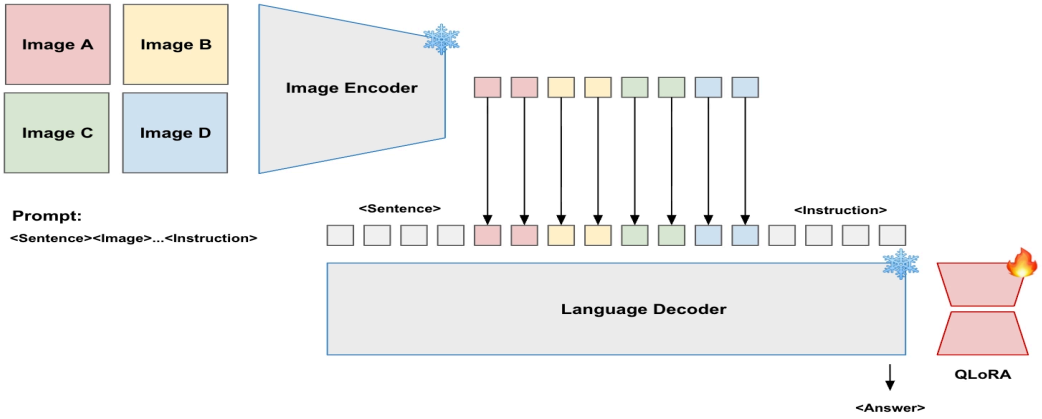
모델이 개별 이미지 사이의 시간 관계를 보다 명시적으로 학습하도록 유도함. 동시에 최종 전체 순서를 함께 예측하게 하여, 쌍별 판단이 하나의 일관된 시간 순서로 결합되도록 함.

2-4. Prompting for Reasoning VLM



o Motivation: Reasoning VLM을 사용한 초기 실험에서 **명시적인 reasoning 종료 토큰을 입력해도, 모델이 이후에 추가적인 추론 문장을 생성하는 경향성을 관찰함.** 또한, reasoning이 포함된 상태로 지도 미세조정을 수행하면 학습 손실은 감소하더라도 평가 단계에서 불필요한 추론 문장을 생성하거나 정답 형식을 준수하지 않는 사례가 빈번히 발생함. 이는 reasoning에 특화된 모델이 사전학습 및 후속 학습 과정에서 획득한 생성 패턴이 본 과제의 간결한 GT 형식과 불일치하기 때문으로 판단함.
o Method: 모델의 기존 생성 결과를 분석하여 **reasoning 과정이 종료된 뒤 최종 답변으로 전환할 때 반복적으로 사용하는 어구를 확인하고, 해당 전환 어구를 프롬프트에 미리 제공하여,** 모델의 출력하는 다음 토큰의 분포가 추가 reasoning보다 최종 정답 생성에 집중되도록 유도함.
o Contribution: **별도의 장문 reasoning 과정 없이 모델이 곧바로 정해진 형식의 최종 답변을 생성하도록 유도함.** 이를 통해 최신 Reasoning VLM의 capacity를 활용하고, 출력 길이와 추론 비용을 줄이는 동시에, reasoning 생성으로 인한 형식 오류와 결과 파싱 실패를 방지함.

2-5. QLoRA Fine-tuning



o Motivation: VLM을 full-finetuing 시, 훈련 성능은 빠르게 향상되었지만, **검증 데이터에서의 성능이 함께 향상되지 않거나 감소하는 과적합 현상을 확인함.** 이에 제한된 규모의 데이터로 파라미터를 모두 업데이트하면, 사전학습 과정에서 획득한 일반적인 표현이 불필요하게 변경될 가능성이 있다고 판단함.
o Method: **Language Decoder의 Attention 및 MLP Layer에만 LoRA Adapter를 적용함.** 구체적으로는 모델의 원본 가중치를 4-bit로 양자화하고, 양자화된 가중치를 고정한 상태에서 LoRA Adapter만 학습하는 **QLoRA 방식을 적용함.**
o Contribution: **사전학습된 VLM의 일반적인 표현 능력을 최대한 보존하며** 시간 순서 판단과 정답 형식 생성에 필요한 Language Decoder의 동작만 효율적으로 조정함. 학습 가능한 파라미터 수를 줄여 **과적합 가능성과 학습 비용을 동시에 줄임.**

항목	값
학습 가능 파라미터	466,911,232 / 27,823,639,792 (1.68%)
LoRA rank	r = 64
적용 대상	언어 디코더 attention + MLP / 비전 인코더 동결

정밀도	4-bit (nf4) + bf16 어댑터

3. Experimental Results

3-1. Ablation Study: Pairwise Temporal Precedence Prediction

Ablation Study: Pairwise Temporal Precedence Prediction

Method	Public	Δ Public	Private	Δ Private
Baseline	0.90226	–	0.85772	–
+ Pairwise Temporal Precedence Prediction	**0.90750**	+0.58%	**0.89024**	+3.79%

Baseline denotes the model trained using standard SFT.

기본 Global Temporal Order Prediction 학습 방식에 **Pairwise Temporal Precedence Prediction을** 추가한 결과, Public score는 상대적으로 **0.58%**, Private score는 **3.79% 향상됨**. 이는 전체 이미지 순서를 직접 예측하는 SFT 학습에 **이미지 쌍별 시간적 선후 관계**를 추가로 학습시키는 방식이 성능 향상에 기여했음을 보여줌.

3-2. Ablation Study: Model Structure

Ablation Study: Model Structure

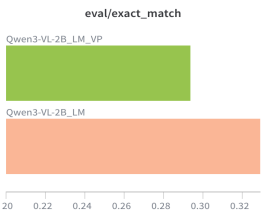
Model	Public	Δ Public	Private	Δ Private
Qwen3.6-27B	**0.94240**	–	**0.91463**	–
Qwen3-VL-32B-Instruct	0.90226	-4.26%	0.85772	-6.22%
Gemma-4-31B-it	0.88307	-6.30%	0.89430	-2.22%

Δ denotes the relative score difference compared with Qwen3.6-27B.

동일한 학습 방법을 적용하여 모델별 성능을 비교한 결과, **Qwen3.6-27B가 가장 높은 성능**을 기록함.
Qwen3-VL-32B-Instruct는 Qwen3.6-27B 대비 Public score는 4.26%, Private score는 6.22% 낮았으며, Gemma-4-31B-it는 Qwen3.6-27B 대비 각각 6.30%, 2.22% 낮은 성능을 보임. 이를 통해 Qwen3.6-27B가 **이미지 간 시간적 관계**를 가장 효과적으로 학습한 모델임을 확인함.

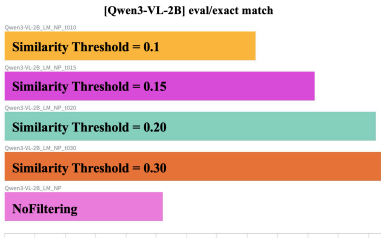
3-3. Ablation Study: Visual Prompting

o Visual Prompting: 각 샘플 이미지에 직접 입력 순서대로 Image 1~ 4 와 같은 텍스트를 삽입함으로써 각 이미지 안에 직접 frame index를 픽셀 레벨 visual index로 렌더링 하여 temporal reasoning을 강화하는 전략임. (CVPR'26)



Qwen3-VL-2B-Instruct 모델에 적용한 결과, validation exact match는 약 0.33에서 오히려 0.29로 감소하여 **상대적으로 약 11%의 성능 하락을 야기함**. 이는 이미지 내부에 입력 순서를 나타내는 시각적으로 추가하는 방식이 shuffle된 시퀀스에서 모델이 순서를 구분해야 하는 태스크에는 효과적이지 않음을 보임. 특히 원본 이미지의 일부를 가리거나 사전학습 당시의 이미지 분포를 변화시켜 시각적 특징 추출을 방해했을 가능성을 시사함.
따라서 **Visual Prompting의 효과가 일관적이지 않다고 간주하고 최종 모델에 적용하지 않음.**

3-4. Ablation Study: Cosine Similarity Threshold in Relevance Filtering



학습 데이터셋에서 코사인 유사도를 기반으로 시각-언어 관련성이 낮은 샘플을 필터링한 경우가, **전체 데이터셋을 사용했을 때 (마지막 행) 보다 검증 결과가 일관적으로 우수함.** 이를 통해 대형 모델의 다운스트림 학습을 위해 고품질의 학습 신호를 선별하는 전략의 유효성을 확인함.
최적의 필터링 조건을 적용하면서 학습 가능한 데이터 개수를 확보하기 위한 임계값으로 0.2 가 적정하다고 판단함.

3-5. Ablation Study: LoRA Module

Ablation Study: Image Encoder Freeze

Method	Public	Δ Public	Private	Δ Private
Baseline	0.89179	–	0.89024	–
+ Image Encoder Freeze	**0.90226**	**+1.17%**	**0.90650**	**+1.83%**

Image Encoder를 고정한 결과, Baseline 대비 Public score는 1.17%, Private score는 1.83% 향상됨. 이는 사전학습된 시각 표현을 유지하고 **Language Decoder를 중심으로 학습하는 방식이 두 평가 데이터 모두에서 성능 향상에 기여했음을 보여줌.**

3-6. Responses in Reasoning VLM (Qwen3.6-27B)

사전학습된 Reasining VLM에서 Assistant Prefix에 따라 모델의 reasoning 여부가 달라지는 것을 분석함.

Assistant Prefix	Sample	Response
<think>\n (default)	1	The user wants me to order four images based on a provided sentence describing a sequence …
	2	The user wants me to order four images based on a provided sentence describing a sequence …
<think>\n\n</think>	1	Let's break down the sentence and map it to the images:\n\nSentence …
	2	[NoNoNoNoNoNo]3214
<think>\nWe need response directly.</think>	1	[NoNoNoYesYesYes][2413]
	2	[NoNoYesNoNoNo][3214]

기본 설정인 "<think>\n"에서는 모든 샘플이 reasoning을 수행했지만, 빈 reasoning 구간인 "<think>\n\n</think>"은 reasoning을 안정적으로 억제하지 못해 샘플별로 서로 다른 응답이 나타남. 반면, **<think> 내부에 직접 응답하라는 명시적 지시를 제공하면 모든 샘플이 reasoning 없이 곧바로 지시한 형태로 정답을 출력함.** 따라서 단순히 <think> 태그를 닫는 것보다, 자연어로 응답 방식을 명시하는 것이 reasoning 제어에 더 효과적인 것으로 판단하고 해당 방식을 채택함.

4. Strengths & Limitations

4-1. Strengths

o 이미지-문장 유사도 기반 필터링으로 학습 데이터의 품질을 향상시키고, Temporal Order Augmentation으로 입력 위치 편향과 정답 클래스 불균형을 줄일 수 있음.
o Global Temporal Order Prediction에 Pairwise Temporal Precedence Prediction을 추가하여, 하나의 전체 순열 정답만을 사용하는 방식보다 조밀하고 세분화된 학습 신호를 제공함.
o Reasoning 모델에서 Reasoning Trajectory를 건너뛰고 바로 정답을 답변할 수 있는 프롬프트 구조를 찾아 활용함으로써 학습 안정성 및 출력 형식 오류를 줄일 수 있음.
o SFT를 활용해 학습을 진행하여 RL 방법과 비교하였을 때 효율적임. 또한, Unsloth 프레임워크와 QLoRA를 함께 적용하여 효율적으로 학습 및 추론이 가능함.

4-2. Limitations

o Reasoning VLM에서 바로 정답을 출력하도록 유도하므로, 복잡한 샘플에서 단계별 reasoning의 이점을 활용하지 못할 수 있음.
o Autoregressive 알고리즘의 한계로 초반에 잘못된 토큰이나 판단을 생성하면 이후 결과에도 영향을 미치며, 이미 생성한 답변을 수정하기 어려움.