VISUAL RAG:

Benchmarking Text-to-Image Retrieval Augmented Generation for Visual Knowledge Intensive Queries

2025 연구실 세미나

Jiyun Kang

Text knowledge vs Visual knowledge

"모나리자는 누구의 작품인가요?"

Text knowledge vs Visual knowledge

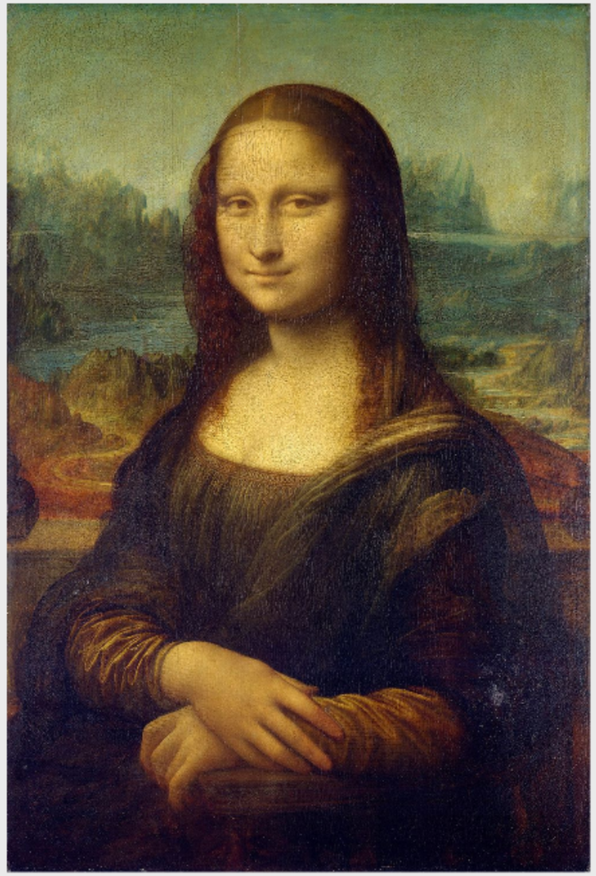
"모나리자는 누구의 작품인가요?"

“모나리자의 배경은 무슨 지형인가요?”

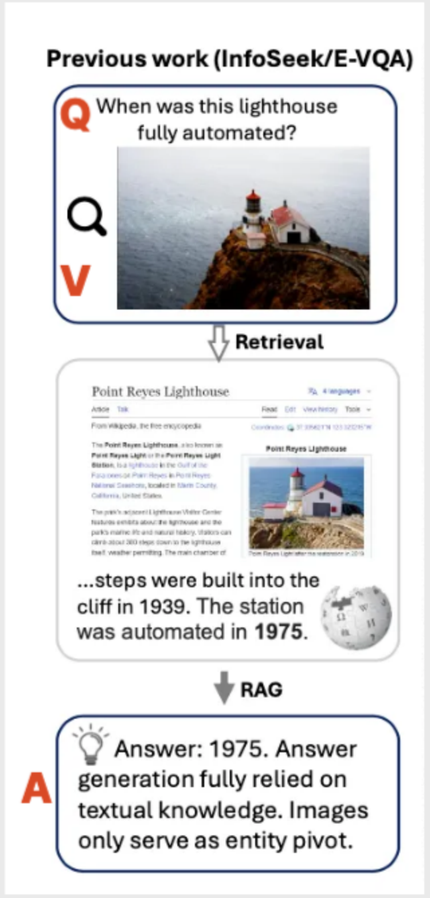
Text knowledge vs Visual knowledge

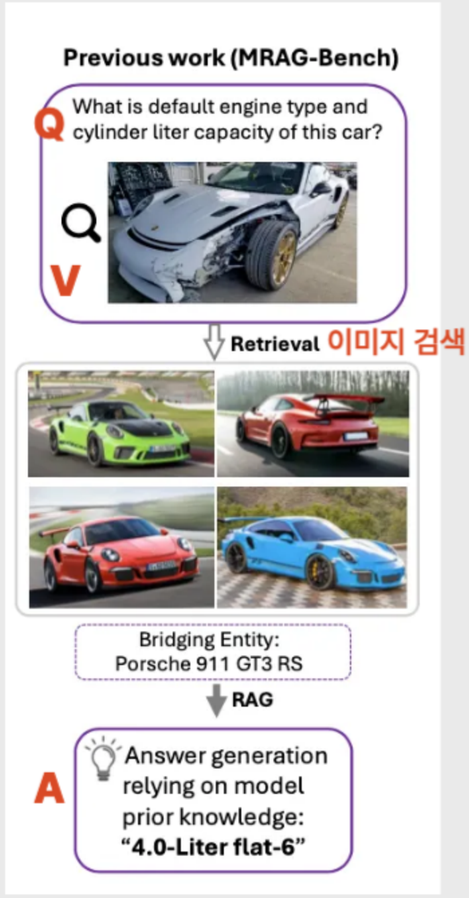
"모나리자는 누구의 작품인가요?"

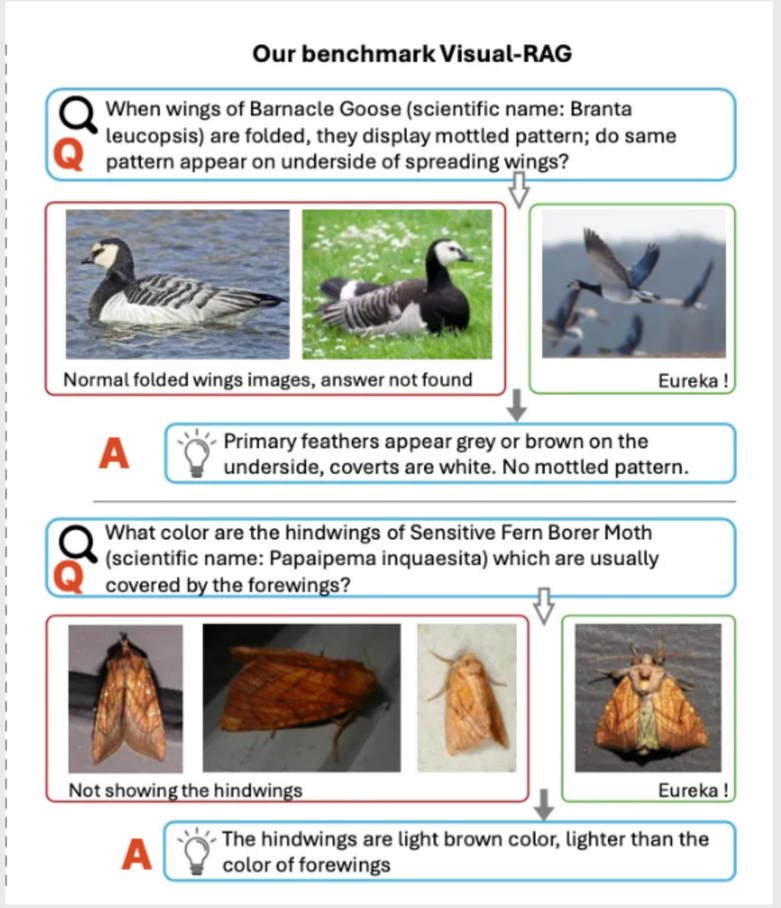
“모나리자의 배경은 무슨 지형인가요?”



VQA Benchmarks (Textual Knowledge Centric) VS QA Benchmarks (Visual Knowledge Centric)



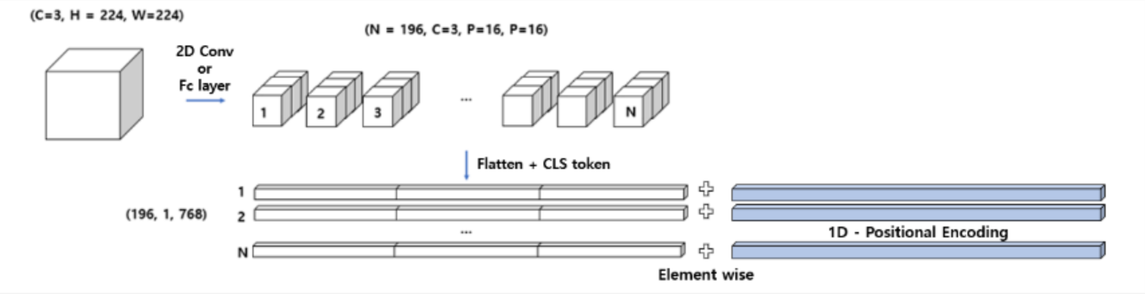


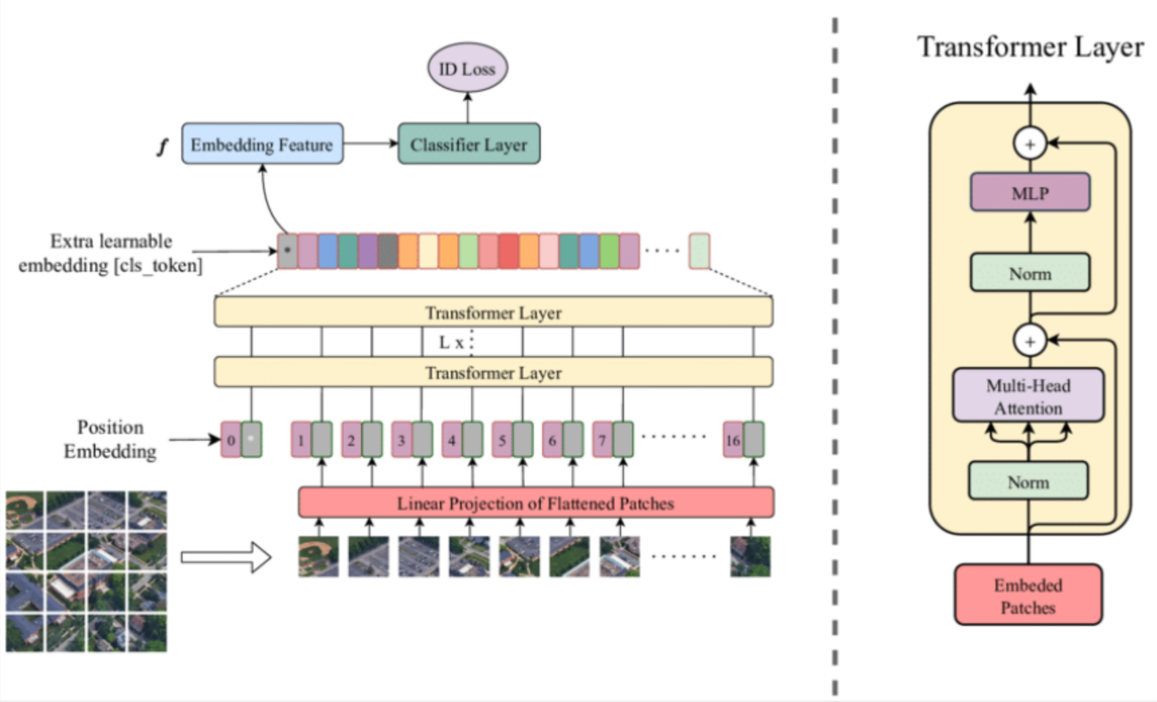


모델 내부 지식에 정답 존재

이미지 자체에 정답 존재

Transformer ViT 의 출력 구조





Retrieval 단계

빠른 검색을 위해 CLS 토큰(글로벌 임베딩)만 사용
> 이미지와 텍스트 임베딩 간 cosine Similarity 계산

$$\text{sim}(f_{text}(q), f_{image}(I)) = \frac{v_q \cdot v_I}{\|v_q\|\|v_I\|}$$

Generation 단계

검색된 이미지에 대해 LLM이 reasoning할 때는 패치 토큰 전체 (혹은 object-level 토큰)를 넘겨줌

- 예: BLIP-2의 Q-Former 구조처럼 LLM이 시각적 토큰 전체에 attention을 걸 수 있도록 함.
- LLM의 입력은

$$\text{Input} = [\text{TEXT}, Z_1^{(L)}, Z_2^{(L)}, ..., Z_N^{(L)}]$$

Benchmark Overview

Task Definition

Q: 텍스트 쿼리(text queries) 집합
D: 이미지 코퍼스 (image corpus)

Visual Knowledge–Intensive Queries (시각 지식 중심 쿼리)

- 모든 쿼리는 텍스트 기반(text-only)
- 질문은 엔티티(category-level) 의 시각적 특성(fine-grained visual feature)에 초점
- 도메인 선택: 생물(organism) 영역

When wings of Barnacle Goose (scientific name: Branta leucopsis) are folded, they display mottled pattern; do same pattern appear on underside of spreading wings?

Text ↔ Image
cross-modal knowledge extraction

Dataset (OVEN (Hu et al., 2023) iNaturalist 2021 (Van Horn et al., 2021))
→ Naturally Co-occurring Hard Negatives

- hard negative: 같은 종(species)의 이미지이지만, 질문된 시각적 특징(feature) 이 나타나지 않는 경우
- 일부러 가짜 이미지나 텍스트로 생성한 게 아닌 자연적으로 발생(naturally co-occurring)

clue image
$\mathcal{I}_c \subset \mathcal{D}$

Distractors(queried visual feature x)
$\mathcal{I}_d = \mathcal{D} \setminus \mathcal{I}_c$

Evaluation Points

(1) visual evidence–centric text-to-image retrieval(시각적 근거 중심의 텍스트-이미지 검색)

→ Retrieval 단계
주어진 질문 q 에 대해, 시스템은 이미지 코퍼스 D에서 상위 k개의 이미지를 검색

$$\mathcal{I}_{ret;k} \subset \mathcal{D}$$

(2) MLLM's ability for visual knowledge-intensive reasoning with retrieved images in retrieval-augmented answer generation
(Retrieval-기반 생성 과정에서의 MLLM의 시각적 추론 능력)

→ Generation 단계
LLM 모델은 q와 상위 k개의 이미지를 입력받아 답을 출력

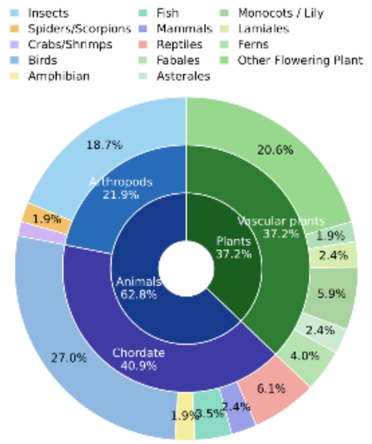
$$(q, I_{ret;k})$$

Core Assumption

"올바른 clue 이미지만 retrieval 되면,
정답은 그 이미지 하나만 보고 직접적으로 추론 가능하다."

Benchmark Construction

Dataset (OVEN (Hu et al., 2023) iNaturalist 2021 (Van Horn et al., 2021))



(a) Organism categories.

전체 이미지 수 (D) 99,017장
평균 clue 이미지 수(per query) 16.19장
Clue 이미지 비율 전체의 6.12% (≈ 1/16)

1. Question generation

LLM(OpenAI-o1)을 이용해 후보 질문 생성

: 한 종(200-300장의 이미지)의 희귀한 특성만을 질문할 수 있도록 질문을 생성

1. 위키피디아 summary를 주고,
이 Passage에 없는 설명들 중 시각적 속성에 관한 질문을 하도록 함.

2. 종 선택의 편향(Biasing for Underrepresented Species)

: 텍스트 검색만으로 답을 쉽게 찾을 수 없게 하기 위해, 위키백과에 덜 등장하는 종 (under-represented species) 위주로 선택

3. 사람이 직접 검수 (문법적 오류, 실제 시각적 속성을 잘 가르키는지 등)

1) 질문 후보들
2) 해당 질문이 다루는 시각적 속성 (ex. hindwing color)

2. MLLM-based coarse filtering

: 생성된 질문들 중에서도, 실제 이미지 데이터 기준으로 정말 희귀한 시각 단서(clue)만 포함하는 질문만 남기기 위해 필터링

$$0 < \rho_q \leq 0.25.$$

1. clue 이미지가 해당 종 이미지 중 소수 (최대 25%) 여야 함
2. 적어도 한 장의 clue 이미지가 존재해야 함

오픈소스 MLLM 1차 필터링 (Llama-3.2-11B-Vision-Instruct)

이미지마다 해당 속성이 "보인다 / 안 보인다"를 이진 분류(binary classification)수행.
(너무 흔한 속성들을 제거)

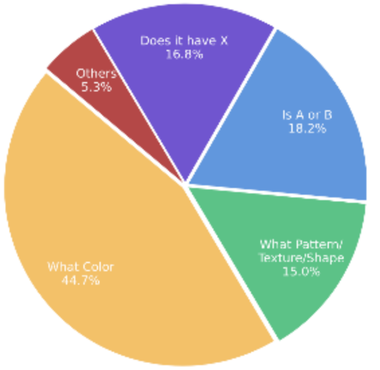
$$\hat{\rho}_q > 0.25.$$

약 1,000개 질문 통과

3. Human Verification

대학생 자원봉사자(annotators)가 직접 이미지 확인

(a) 각 이미지에서 시각 특징이 보이는지 여부 (labeling)
(b) 질문에 대한 정답(answer) 제공.



(b) Question categories.

374개 질문 최종 통과
(clue rate 조건(0 < ρ ≤ 0.25) 만족 및 정답 가능 확인)

Experiments Setup - Evaluation

01

RETRIEVAL EVALUATION (이미지 검색 평가)

$\mathcal{I}_{ret;k} \subset \mathcal{D}$

주어진 텍스트 질문에 대해, 모델이 올바른 clue 이미지를 얼마나 잘 찾아내는지 평가

02

GENERATION EVALUATION (답변 생성 평가)

$(q, I_{ret;k})$

검색된 이미지를 바탕으로, MLLM이 얼마나 정확하고 사실적인 답을 생성하는지 평가

Experiments Setup - Evaluation

평가 지표

01

RETRIEVAL EVALUATION
(이미지 검색 평가)

$\mathcal{I}_{ret;k} \subset \mathcal{D}$

- **RECALL@K**
  - **검색의 포괄성(coverage) 측정**
  - 전체 clue 이미지 중 상위 k 안에 포함된 비율
- **NDCG@K**
  - **정확한 순위 품질(ranking quality) 평가**
  - (Normalized Discounted Cumulative Gain)
    순위가 높은 위치에 clue 이미지가 있을수록 높은 점수 부여
- **HIT@K**
  - **단일 성공 여부(success/fail) 측정**
  - top-k 안에 최소 1개의 clue 이미지가 존재하면 1, 없으면 0
- **HIT COUNT@K**
  - **단서 수(evidence quantity) 측정**
  - top-k 안에 포함된 clue 이미지의 개수
    (“모델이 유용한 근거를 얼마나 많이 공급했는가”)

Experiments Setup - Evaluation

01

RETRIEVAL EVALUATION (이미지 검색 평가)

$\mathcal{I}_{ret;k} \subset \mathcal{D}$

검색 범위

global retrieval -> species-specific retrieval

query가 구체적이다보니 전체 데이터셋에서 찾기는 어렵다고 판단
-> 검색 범위를 해당 종(species) 의 이미지 집합으로 좁혔음.

Search Index

FAISS Library using flat inner product index

모든 이미지와 질의를 embedding 벡터로 변환한 뒤,
내적 유사도(cosine similarity)에 따라 top-k 이미지를 반환

Dense Retriever

1) CLIP
clip-vit-large-patch14-336 (Radford et al., 2021)

2) BGE-VL-large
CLIP의 향상 버전 (Zhou et al., 2024)

3) VLM2Vec-qwen2VL-2B
MLLM 기반 retriever (Jiang et al., 2025)

Experiments Setup - Evaluation

02

GENERATION EVALUATION (답변 생성 평가)

$(q, I_{ret;k})$

기존 방식

exact-match or recall-based metrics

부분적으로 비슷한 애들은 틀렸음에도 불구하고 정답으로 처리해버리는 문제 발생
ex. “black and white” 가 정답인데,
“black and white with yellow dots”도 정답으로 처리

새로운 평가 방식

LLM-as-Judge (OpenAI-o4-mini)

“이 답이 맞는지 / 틀린지 / 부분적으로 맞는지”
annotator(사람)이 제공한 여러 정답(ground-truth candidates)을 모두 허용 범위로 사용
지시 프롬프트(instruction prompt) : LLM에게 “다른 표현으로 말한 동일 의미도 인정하라”고 명시

Format Instructions: Separate the remarks with score using “|”, that is, use the syntax of: “Score: score | Likely Hallucination”, “Score: score”, “Score: score | Likely Hallucination | Redundant”, “Score: 0 | No Answer ”. If any explanation on why giving the score is needed, do not start a new line and append after remark with brackets, e.g. “Score: score | Redundant | (Explanation: abc)”.

Following are few examples:

Question: Is there any specific color marking around the eyes of a semipalmated plover (scientific name: Charadrius semipalmatus)?
Reference Answer: black eye-round feather, white stripe above eyes. (sometimes connected to the white forehead)

Student Answer: Yes, the bird has a distinctive black line that runs through the eye, which is a key identifying feature.
Score: 0 | Likely Hallucination

Student Answer: They have a black vertical band in front of the eye, a white band above the eye, and a single black band that wraps partially around the eye, creating a partial “mask” appearance.
Score: 1
Student Answer: Yes, the semipalmated plover has a distinctive black/dark ring around its eye, surrounded by a bright white ring or patch
Score: 0.5 | Likely Hallucination (Explanation: not white ring, but only a line above the eye)

Question: What is the typical color of the antennae of Harris’s checkerspot butterfly (scientific name: Chlosyne harrisii)?
Reference Answer: alternating black and white band, with yellow on the tip

Student Answer: The antennae of Harris’s checkerspot butterfly are black with orange-tipped clubs.
Score: 0.5 (Explanation: not mentioning black and white)

Experiments Setup - Evaluation

02

GENERATION EVALUATION
(답변 생성 평가)

$(q, I_{ret;k})$

기존 방식

exact-match or recall-based metrics

부분적으로 비슷한 애들은 틀렸음에도 불구하고 정답으로 처리해버리는 문제 발생
ex. “black and white” 가 정답인데,
“black and white with yellow dots”도 정답으로 처리

새로운 평가 방식

LLM-as-Judge (OpenAI-o4-mini)

신뢰도 검증 결과
200개의 샘플을 사람과 LLM이 동시에 평가한 결과를 비교 (F1 = 0.92)

Human\LLM	True	Partial	False
True	86	1	3
Partial	6	14	2
False	6	17	65

Experiments Setup - Multimodal LLMs

Proprietary Models (상용 모델)

- GPT-4o (OpenAI 2025)
- Gemini-2.5-Pro (Google 2025)
- Claude-Sonnet-4 (Anthropic 2025)

Open-Source Models (오픈소스 모델) - 5B~12B

- Phi-4-Multimodal-Instruct (5.6B, Microsoft 2025)
- Qwen2.5-VL-7B-Instruct (Bai et al. 2025)
- InternVL-3-8B (Zhu et al. 2025)
- Llama-3.2-11B-Vision-Instruct (Meta Llama Team 2024)
- Pixtral-12B (Mistral AI 2024)

- 모든 모델은 “multi-image input 지원” 을 기본 전제로 사용됨.

- 모델 크기(파라미터 수)는 5B~12B까지 다양
→ 모델 규모에 따른 fine-grained 시각 추론 성능 차이를 비교 가능.

- 추론 세팅(Inference Parameters) 은 각 모델의 기본값(default parameters)을 그대로 사용.

Result - Text-to-Image Retrieval

전체 clue 이미지 중 상위 k 안에 포함된 비율
순위가 높은 위치에 clue 이미지가 있으면 높은 점수 부여
top-k 안에 최소 1개의 clue 이미지가 존재하면 1, 없으면 0
top-k 안에 포함된 clue 이미지의 개수

CLIP	@1	@5	@10	@20	@30
Recall	2.81	10.26	16.70	25.54	33.16
NDCG	24.33	25.97	30.98	39.21	46.25
Hit	24.33	53.74	67.65	77.54	82.62
Hit Count	0.24	1.01	1.84	3.18	4.38
BGE	@1	@5	@10	@20	@30
Recall	3.63	12.43	18.81	28.64	35.97
NDCG	26.74	29.15	33.4	41.83	48.84
Hit	26.74	58.56	68.98	79.14	83.96
Hit Count	0.27	1.17	2.04	3.43	4.60
VLM2Vec	@1	@5	@10	@20	@30
Recall	2.91	10.01	16.03	26.37	33.84
NDCG	24.87	26.33	30.09	39.13	45.93
Hit	24.87	51.87	62.83	78.34	82.35
Hit Count	0.25	1.10	1.91	3.33	4.49

Table 2: Retrieval results using the sub-corpus of 200-300 images for each query. Even within the small, species level corpus, the model struggles with our challenging text-to-image retrieval task.

Overall difficulty

clue 이미지를 못 찾으면, MLLM이 아무리 똑똑해도 잘못된 이미지들을 보고 답을 생성해야 하므로 전체 성능이 제한

Results

(1) clue images are genuinely rare
(2) 단서를 찾아도 retriever가 similarity 기준으로 높은 순위에 올리지 못함

Takeaway

cross-modal retriever으로는 성능을 높이는 것이 어려움.
K=3으로 늘려도 여전히 성능이 크게 올라가지 않음.
→ retriever 자체의 질적 향상이 필요함.

Result :
Visual Retrieval-Augmented Generation

1 Does image as augmentation benefit MLLMs? (“이미지가 정말 도움이 되나?”)

조건	설명	목적
Zero-shot	이미지 없이 텍스트 질문만 입력	baseline (비교 기준)
GT clue (Ground Truth Clue)	실제 정답 단서(clue)가 들어 있는 이미지 1장 제공	upper bound (최대 효과)
Non-clue	같은 종(species)의 무관한 이미지 1장 제공	“이미지가 단순히 있어서 생긴 효과인가?” 검증

- 시각적 단서가 포함된 이미지가 있어야 성능이 향상됨.
- open source model은 +15이상 평균이 향상됨.
- 상용 모델은 평균이 원래 높은데, 약 +2~6점 향상됨
- Non-clue는 오히려 성능을 하락시킴.

Model	Phi4-MM	Qwen2.5VL	InternVL3	Pixtral	Llama3.2-V	GPT-4o	Gemini	Claude
Baselines								
Zero-shot (no image)	35.16	38.90	39.17	41.71	32.35	53.74	60.43	54.28
GT clue (1 image)	45.04	41.79	43.69	47.11	47.81	59.81	62.88	56.79
Non-clue (1 image)	34.76	30.08	29.28	42.74	40.37	14.97	17.11	21.39
Top-K Retrieval-Augmented Generation								
k=1	39.17	41.98	37.57	42.11	44.25	24.06	32.22	35.03
3	39.30	44.39	41.18	39.71	**46.79**	41.18	48.53	45.32
5	37.70	46.93	39.97	41.71	43.80	47.86	54.95	50.40
7	**41.44**	45.99	**42.65**	41.44	43.98	51.34	55.35	52.14
10	**41.44**	48.26	42.11	42.65	43.42	49.73	57.62	53.88
15	41.04	49.73	41.04	**44.39**	44.92	50.80	60.03	55.48
20	41.31	**50.53**	41.31	-	-	**52.67**	**61.50**	**57.35**
One-in-K Augmented Generation								
k=3	**41.92**	46.85	**41.06**	**47.11**	**46.73**	48.95	59.57	48.05
5	40.13	48.04	38.44	46.68	45.19	53.22	60.70	49.86
7	40.80	47.78	39.76	46.68	44.20	55.36	61.53	51.01
10	39.08	47.43	38.50	44.83	41.90	56.25	62.70	51.69
15	40.40	**49.57**	39.14	43.95	41.12	56.36	**64.01**	**52.67**
20	40.83	48.55	39.09	-	-	**56.50**	63.47	52.14

Table 3: Main experiment results. The coloured cells shows the difference with zero-shot score, pink cells indicate performance under zero-shot baseline, light green cells indicate performance over zero-shot, but lower than GT, and green cells indicate outperforming GT. All models benefit from ground-truth clue image as augmentation.

Result :
Visual Retrieval-Augmented Generation

2 Does realistic visual RAG benefit MLLMs?
("현실적인 RAG 상황에서도 여전히 도움이 되나?")

조건	설명	목적
Qwen2.5-VL & Proprietary models (GPT-4o, Gemini, Claude)	k가 커질수록 성능 향상	clue 이미지가 포함되면 잘 활용함
나머지 Open-source MLLM들 (Phi-4, InternVL, Llama-3.2, Pixtral)	k=5~10에서 성능 포화	이후엔 거의 향상 없음, 오히려 약간 하락

→ 즉, top-k 를 높이면 오히려 Noise가 발생해서 혼란스러운 경우도 발생할 수 있음.

+ Multi-image 처리 한계
→ 대부분 오픈소스 MLLM은 한 번에 많은 이미지를 효율적으로 통합(attend)하지 못함

Model	Phi4-MM	Qwen2.5VL	InternVL3	Pixtral	Llama3.2-V	GPT-4o	Gemini	Claude
Baselines								
Zero-shot (no image)	35.16	38.90	39.17	41.71	32.35	53.74	60.43	54.28
GT clue (1 image)	45.04	41.79	43.69	47.11	47.81	59.81	62.88	56.79
Non-clue (1 image)	34.76	30.08	29.28	42.74	40.37	14.97	17.11	21.39
Top-K Retrieval-Augmented Generation								
k=1	39.17	41.98	37.57	42.11	44.25	24.06	32.22	35.03
3	39.30	44.39	41.18	39.71	**46.79**	41.18	48.53	45.32
5	37.70	46.93	39.97	41.71	43.80	47.86	54.95	50.40
7	**41.44**	45.99	**42.65**	41.44	43.98	51.34	55.35	52.14
10	**41.44**	48.26	42.11	42.65	43.42	49.73	57.62	53.88
15	41.04	49.73	41.04	**44.39**	44.92	50.80	60.03	55.48
20	41.31	**50.53**	41.31	-	-	**52.67**	**61.50**	**57.35**
One-in-K Augmented Generation								
k=3	**41.92**	46.85	**41.06**	**47.11**	**46.73**	48.95	59.57	48.05
5	40.13	48.04	38.44	46.68	45.19	53.22	60.70	49.86
7	40.80	47.78	39.76	46.68	44.20	55.36	61.53	51.01
10	39.08	47.43	38.50	44.83	41.90	56.25	62.70	51.69
15	40.40	**49.57**	39.14	43.95	41.12	56.36	**64.01**	**52.67**
20	40.83	48.55	39.09	-	-	**56.50**	63.47	52.14

Table 3: Main experiment results. The coloured cells shows the difference with zero-shot score, pink cells indicate performance under zero-shot baseline, light green cells indicate performance over zero-shot, but lower than GT, and green cells indicate outperforming GT. All models benefit from ground-truth clue image as augmentation.

Result :
Visual Retrieval-Augmented Generation

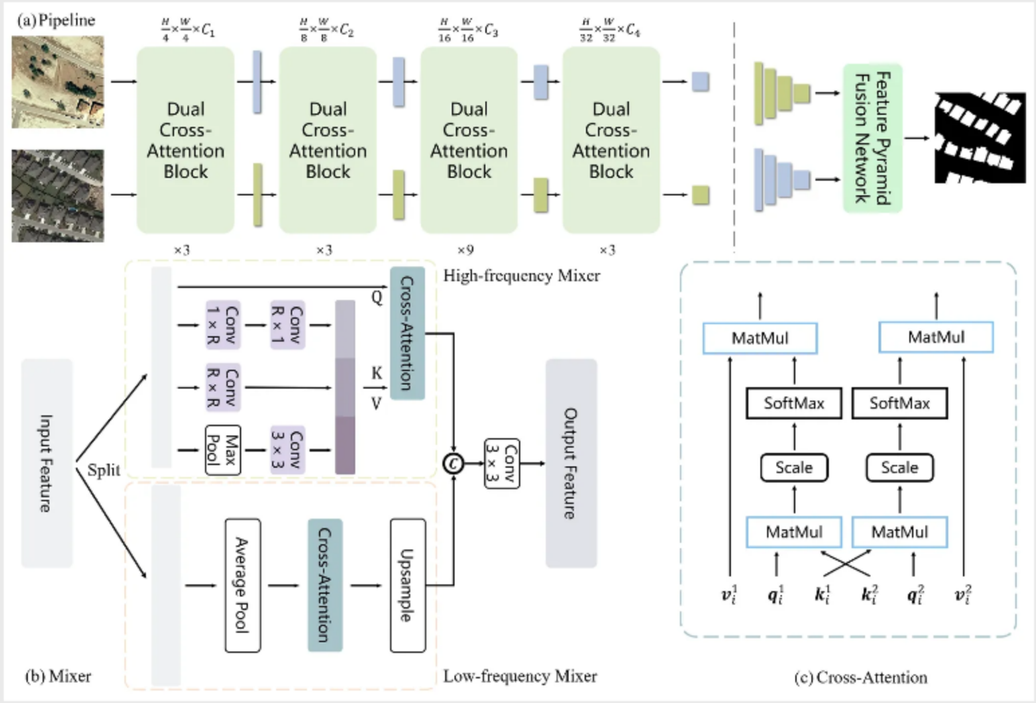
왜 Open-source MLLMS에서 Multi-image를 처리할 수 없는가?

(1) Transformer 구조의 한계
- MLLM(예: LLaVA, Qwen-VL, InternVL)은 대부분 Vision Transformer (ViT)를 사용
- ViT는 입력 이미지를 "패치 토큰"으로 나누고, 각 이미지를 독립적으로 인코딩한 뒤 CLS 토큰을 요약 벡터로 사용.

하지만 여러 장의 이미지를 입력받으면,
- 각 이미지의 패치 토큰 수 × 이미지 개수 → 토큰 수 폭발(token explosion)
- 예: 한 이미지당 256개 패치 × 10장 = 2560개 비전 토큰
- → 연산량이 급격히 증가하고 attention 비용이 폭발 ($O(N^2)$)

(2) Cross-image Attention 부재
- GPT-4o나 Gemini-2.5는 **multi-image cross-attention layer** 를 갖고 있어 이미지 간 관계를 직접 학습

- 반면 오픈소스 모델(예: LLaVA, Phi-4, InternVL 등)은 대부분 이미지들을 독립적으로 인코딩 후, 단순 평균(mean pooling)해서 텍스트 인풋



Result :
Visual Retrieval-Augmented Generation

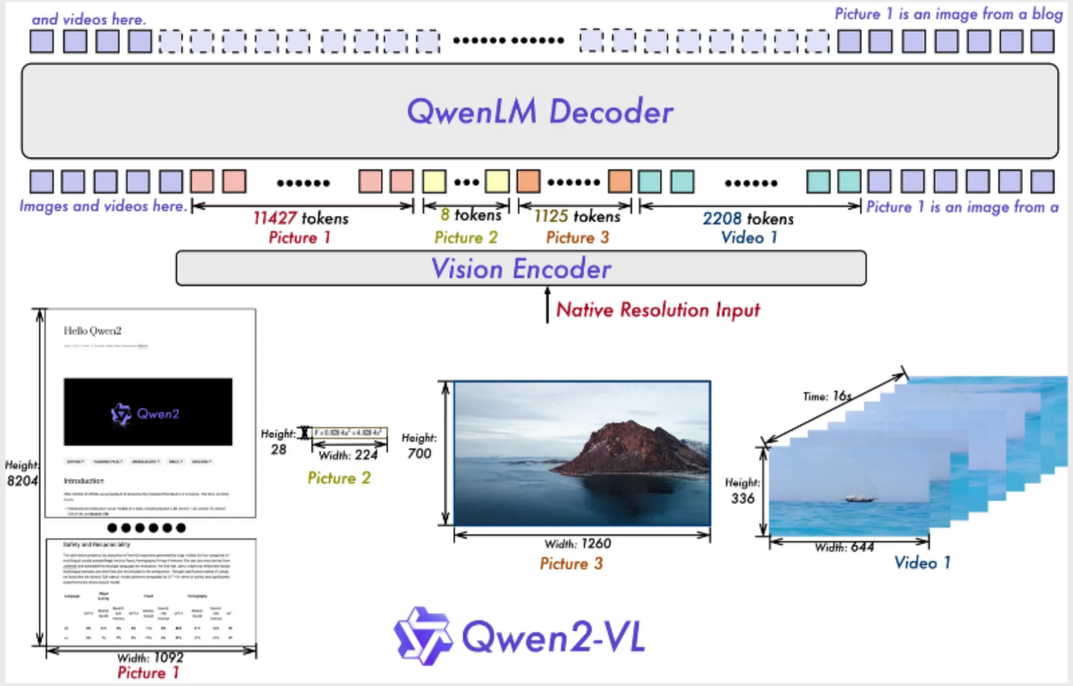
QWEN2.5 VL은 어떤 모델이길래 이미지 간 관계를 파악할 수 있을까?

(1) Image Patch Encoding (비전 인코더)
각 입력 이미지는 Vision Transformer (ViT) 기반의 인코더로 처리
- 이미지를 일정 크기의 패치(patch)로 나누고, 각 패치가 비전 토큰(visual token) 으로 변환
- ex. Image 1 → [CLS_1, v_{11}, v_{12}, ... v_1N]

(2) Token Alignment (시각 토큰 정렬 및 융합)
- 단순히 이어붙이는(concatenation) 게 아니라, 각 이미지의 CLS 토큰에 "이미지 위치 인코딩(image index embedding)" 을 추가하여 모델이 "이 토큰이 어떤 이미지에서 왔는지"를 인식
- ex. [CLS_1, v_{11}, v_{12}, ... v_1N, CLS_2, v_{21}, ... v_2N, CLS_3, ...]

(3) Cross-Attention Layers (이미지 간 관계 학습 핵심)
Query: 텍스트 토큰 (질문)
Key/Value: 모든 이미지의 비전 토큰
-> 하나의 텍스트 질의가 모든 이미지의 시각적 특징들과 동시에 교차-attention을 수행

$$Attention(Q_{text}, K_{visual\_1,visual\_2,...}, V_{visual\_1,visual\_2,...})$$



Result :
Visual Retrieval-Augmented Generation

3 Does a guaranteed clue outweigh the noise of additional non-clue-images?
(단서 이미지가 확실히 포함되어 있더라도, 추가된 비단서 (non-clue) 이미지들의 노이즈를 이길 수 있을까?")

추가 실험 진행 (1-in-k)

구성	설명
이미지 구성	정확히 1장의 clue 이미지 + (k−1)장의 non-clue 이미지
의도	retriever의 오류(단서를 못 찾음)를 제거하고, 모델 자체의 시각적 처리 한계만 평가
clue 위치	항상 맨 앞(첫 번째 이미지) → "완벽한 retrieval이 이루어진 상황"을 가정

Model	Phi4-MM	Qwen2.5VL	InternVL3	Pixtral	Llama3.2-V	GPT-4o	Gemini	Claude
Baselines								
Zero-shot (no image)	35.16	38.90	39.17	41.71	32.35	53.74	60.43	54.28
GT clue (1 image)	45.04	41.79	43.69	47.11	47.81	59.81	62.88	56.79
Non-clue (1 image)	34.76	30.08	29.28	42.74	40.37	14.97	17.11	21.39
Top-K Retrieval-Augmented Generation								
k=1	39.17	41.98	37.57	42.11	44.25	24.06	32.22	35.03
3	39.30	44.39	41.18	39.71	**46.79**	41.18	48.53	45.32
5	37.70	46.93	39.97	41.71	43.80	47.86	54.95	50.40
7	**41.44**	45.99	**42.65**	41.44	43.98	51.34	55.35	52.14
10	**41.44**	48.26	42.11	42.65	43.42	49.73	57.62	53.88
15	41.04	49.73	41.04	**44.39**	44.92	50.80	60.03	55.48
20	41.31	**50.53**	41.31	-	-	**52.67**	**61.50**	**57.35**
One-in-K Augmented Generation								
k=3	**41.92**	46.85	**41.06**	**47.11**	**46.73**	48.95	59.57	48.05
5	40.13	48.04	38.44	46.68	45.19	53.22	60.70	49.86
7	40.80	47.78	39.76	46.68	44.20	55.36	61.53	51.01
10	39.08	47.43	38.50	44.83	41.90	56.25	62.70	51.69
15	40.40	**49.57**	39.14	43.95	41.12	56.36	**64.01**	**52.67**
20	40.83	48.55	39.09	-	-	**56.50**	63.47	52.14

Table 3: Main experiment results. The coloured cells shows the difference with zero-shot score, pink cells indicate performance under zero-shot baseline, light green cells indicate performance over zero-shot, but lower than GT, and green cells indicate outperforming GT. All models benefit from ground-truth clue image as augmentation.

Result :
Visual Retrieval-Augmented Generation

3 Does a guaranteed clue outweigh the noise of additional non-clue-images?
(단서 이미지가 확실히 포함되어 있더라도, 추가된 비단서 (non-clue) 이미지들의 노이즈를 이길 수 있을까?")

Open-source models

Qwen을 제외한 대부분의 오픈소스 모델들은 k값이 커질수록 정확도가 떨어졌음.
즉, clue가 보장되어 있음에도 불구하고 non-clue 이미지가 많아질수록 혼란스러워함

→ 오픈소스 모델은 retriever의 문제가 아닌 모델 자체의 문제

Model	Phi4-MM	Qwen2.5VL	InternVL3	Pixtral	Llama3.2-V	GPT-4o	Gemini	Claude
Baselines								
Zero-shot (no image)	35.16	38.90	39.17	41.71	32.35	53.74	60.43	54.28
GT clue (1 image)	45.04	41.79	43.69	47.11	47.81	59.81	62.88	56.79
Non-clue (1 image)	34.76	30.08	29.28	42.74	40.37	14.97	17.11	21.39
Top-K Retrieval-Augmented Generation								
k=1	39.17	41.98	37.57	42.11	44.25	24.06	32.22	35.03
3	39.30	44.39	41.18	39.71	**46.79**	41.18	48.53	45.32
5	37.70	46.93	39.97	41.71	43.80	47.86	54.95	50.40
7	**41.44**	45.99	**42.65**	41.44	43.98	51.34	55.35	52.14
10	**41.44**	48.26	42.11	42.65	43.42	49.73	57.62	53.88
15	41.04	49.73	41.04	**44.39**	44.92	50.80	60.03	55.48
20	41.31	**50.53**	41.31	-	-	**52.67**	**61.50**	**57.35**
One-in-K Augmented Generation								
k=3	**41.92**	46.85	**41.06**	**47.11**	**46.73**	48.95	59.57	48.05
5	40.13	48.04	38.44	46.68	45.19	53.22	60.70	49.86
7	40.80	47.78	39.76	46.68	44.20	55.36	61.53	51.01
10	39.08	47.43	38.50	44.83	41.90	56.25	62.70	51.69
15	40.40	**49.57**	39.14	43.95	41.12	56.36	**64.01**	**52.67**
20	40.83	48.55	39.09	-	-	**56.50**	63.47	52.14

Table 3: Main experiment results. The coloured cells shows the difference with zero-shot score, pink cells indicate performance under zero-shot baseline, light green cells indicate performance over zero-shot, but lower than GT, and green cells indicate outperforming GT. All models benefit from ground-truth clue image as augmentation.

Result :
Visual Retrieval-Augmented Generation

3 Does a guaranteed clue outweigh the noise of additional non-clue-images?
(단서 이미지가 확실히 포함되어 있더라도, 추가된 비단서 (non-clue) 이미지들의 노이즈를 이길 수 있을까?")

상용 모델(GPT-4o, Gemini, Claude)

k가 증가함에 따라 성능이 지속적으로 상승했으며, 특히 GPT-4o와 Gemini는 기존 top-k RAG 성능을 넘어섬
(이미지가 많아져도 clue를 잘 식별하고 noise를 거를 수 있는 안정적인 구조임을 보여줌.)

→ 반면, 상용 모델들이 top-k RAG 실험에서 낮은 k(예: k=1~5)에서 보였던 다소 낮은 성능은, clue 이미지를 retrieval 단계에서 못 가져온 것 때문이라는 점이 확인

Model	Phi4-MM	Qwen2.5VL	InternVL3	Pixtral	Llama3.2-V	GPT-4o	Gemini	Claude
Baselines								
Zero-shot (no image)	35.16	38.90	39.17	41.71	32.35	53.74	60.43	54.28
GT clue (1 image)	45.04	41.79	43.69	47.11	47.81	59.81	62.88	56.79
Non-clue (1 image)	34.76	30.08	29.28	42.74	40.37	14.97	17.11	21.39
Top-K Retrieval-Augmented Generation								
k=1	39.17	41.98	37.57	42.11	44.25	24.06	32.22	35.03
3	39.30	44.39	41.18	39.71	**46.79**	41.18	48.53	45.32
5	37.70	46.93	39.97	41.71	43.80	47.86	54.95	50.40
7	**41.44**	45.99	**42.65**	41.44	43.98	51.34	55.35	52.14
10	**41.44**	48.26	42.11	42.65	43.42	49.73	57.62	53.88
15	41.04	49.73	41.04	**44.39**	44.92	50.80	60.03	55.48
20	41.31	**50.53**	41.31	-	-	**52.67**	**61.50**	**57.35**
One-in-K Augmented Generation								
k=3	**41.92**	46.85	**41.06**	**47.11**	**46.73**	48.95	59.57	48.05
5	40.13	48.04	38.44	46.68	45.19	53.22	60.70	49.86
7	40.80	47.78	39.76	46.68	44.20	55.36	61.53	51.01
10	39.08	47.43	38.50	44.83	41.90	56.25	62.70	51.69
15	40.40	**49.57**	39.14	43.95	41.12	56.36	**64.01**	**52.67**
20	40.83	48.55	39.09	-	-	**56.50**	63.47	52.14

Table 3: Main experiment results. The coloured cells shows the difference with zero-shot score, pink cells indicate performance under zero-shot baseline, light green cells indicate performance over zero-shot, but lower than GT, and green cells indicate outperforming GT. All models benefit from ground-truth clue image as augmentation.

Result :
Visual Retrieval-Augmented Generation

How efficiently do MLLMs utilize clue and how robust are they against distractors?

generalized clue utilization efficiency (gCUE) -단서 활용 효율 지표

: 현재 clue 활용 효율을 정규화(normalized) 해서 보여주는 척도

$$\mathrm{gCUE}(k;\lambda) = \frac{A_k - B^{(\lambda)}}{A_{GT} - B^{(\lambda)}}.$$

- 1-in-k 세팅에서의 정확도 (clue 1장 + (k-1) non-clue 이미지)
- GT clue 단일 이미지 실험에서의 정확도

$$B^{(\lambda)} = \lambda A_Z + (1-\lambda) A_{NC}$$

- 기준선(baseline)-이미지가 없을 때(zero-shot) ,
무관한 이미지를 봤을 때(non-clue)의 성능을 가중평균으로 통합한 값
(논문에서는 기본값 0.5, 즉 두 baseline을 동일하게 반영함.)

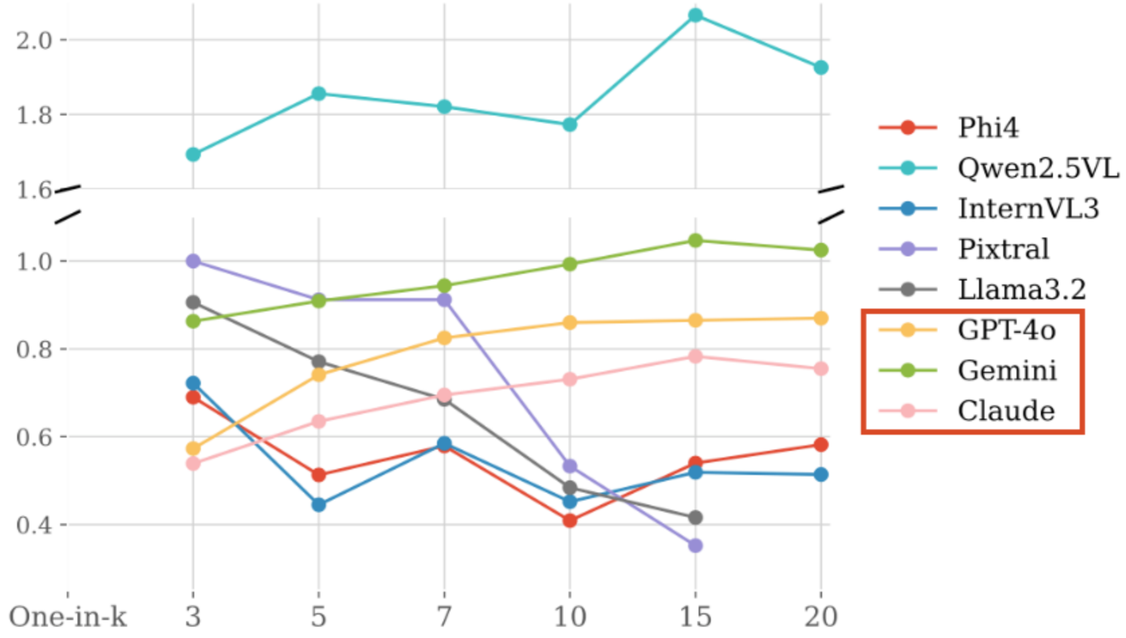


Figure 3: gCUE metrics by models and k.

Result :
Visual Retrieval-Augmented Generation

4

How efficiently do MLLMs utilize clue and how robust are they against distractors?

Qwen2.5VL

- k가 커져도 gCUE가 안정적이거나 오히려 상승 경향.
- 즉, 단서 활용 효율이 높고, noise에 강함.
- clue와 non-clue를 비교·대조하여 clue 신호를 효과적으로 강화

다른 오픈소스 모델들 (Phi-4, InternVL, Llama-3.2, Pixtral 등)

- gCUE가 k 증가에 따라 지속적으로 감소 (decline)
- 즉, 단서의 영향력이 점점 약화됨.

상용 모델들 (GPT-4o, Gemini, Claude)

- 작은 k에서는 낮은 gCUE → 소수의 noise에도 clue 신호가 가려짐.
- 하지만 k가 커질수록 gCUE가 회복·상승 → noise 속에서 clue를 점점 구분함.

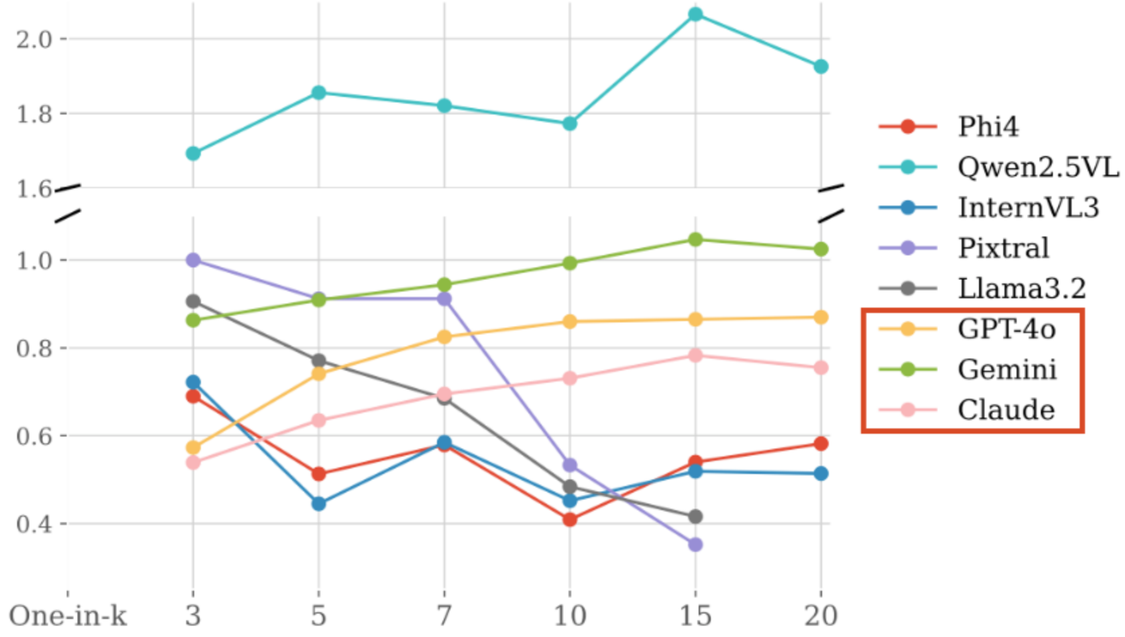


Figure 3: gCUE metrics by models and k.

Results

① 시각적 근거 검색은 어렵다

: cross-modal retriever가 fine-grained 단서 식별 실패

② 단일 clue 이미지는 유효하다

: 정답 이미지만 주면 모델이 잘 추론함

③ 현실적 환경에서는 어렵다

: top-k 검색 시 성능 급감 (여러 이미지 중 단서 활용 능력 부족)

④ 오픈소스 vs 상용 모델의 반대 양상

: open-source는 k↑ → 혼란, proprietary는 k↑ → 안정

THANK YOU