MEDICAL VLM · 주간 세미나 · 논문 3편 리뷰

의료 VLM은 이미지를 보고 있는가?

PAPER ①

Multimodal foundation models exploit text to make medical image predictions

Nature Communications 2026 · Harvard Medical School

PAPER ②

MMedPO: Aligning Medical Vision-Language Models with Clinical-Aware Multimodal Preference Optimization

ICML 2025 · UNC Chapel Hill 외

PAPER ③

Med-CMR: A Fine-Grained Benchmark Integrating Visual Evidence and Clinical Logic

arXiv 2512.00818 · NUS 외

이정섭

NLP&AI

Korea University

오늘 리뷰할 논문 3편

PAPER ①

Exploiting Text

Nature Communications 2026

멀티모달 모델의 의료 영상 정확도는 **실제로 어느 모달리티에서 오는가?**

PAPER ②

MMedPO

ICML 2025

시각 근거를 무시하는 답을 **손실함수로 벌줄 수 있는가?**

PAPER ③

Med-CMR

arXiv 2512.00818

무엇이 얼마나 부족한지를 **능력 축별로 나눠 잴 수 있는가?**

VLM 30초 복습 - 이미지는 어디로 들어가는가

거의 모든 현대 VLM은 **이미지를 토큰으로 바꿔 LLM 입력열 앞에 붙이는** 구조다. 즉 이미지는 LLM에게 "또 하나의 텍스트 조각"처럼 보인다.

이미지

흉부 X-ray, 피부 병변 사진, 병리 슬라이드 ...

→

Vision Encoder

CLIP/SigLIP ViT가 패치 단위 특징 추출

보통 frozen

→

Projector

MLP 또는 Q-Former가 시각 특징을 LLM 임베딩 공간으로 사상 → 이미지 토큰 N개

학습되는 얇은 다리

→

LLM

[이미지 토큰들] + [질문 텍스트]를 하나의 열로 받아 자기회귀 생성

파라미터의 대부분

예고되는 비대칭

- 파라미터·사전학습 토큰의 **절대 다수가 언어 쪽**에 있다
- projector는 보통 수천만 파라미터 규모의 **얇은 다리**일 뿐
- 이미지 토큰은 LLM 입장에서 "낯설고 노이지한 접두부"

질문

- LLM이 이미지 토큰을 **실제로 근거로 삼는가**, 아니면 텍스트 문맥만으로 답을 정하는가?
- 둘이 **충돌**하면 누가 이기는가?
- 이 질문에 숫자로 답한 것이 오늘 첫 논문이다

의료 VLM 계보 - 무엇이 있고 무엇이 다른가

공통점: 대부분 **범용 VLM 구조를 그대로 쓰고 의료 이미지-텍스트 쌍으로 추가 학습**한다. 구조가 아니라 데이터가 "의료"다.

① 범용 VLM (비교군)

GPT-4V / GPT-4o / o3 / GPT-5

OpenAI · 2023 – 2025

의료 전용 학습 없음. 그럼에도 의료 벤치마크에서 강함

Gemini · Claude · Llama 3.2-Vision

2023 – 2025

proprietary / open-weight 범용 계열

② 의료 특화 VLM

LLaVA-Med

Microsoft · NeurIPS 2023 D&B

PMC 그림-캡션으로 curriculum 학습. 오늘 두 번째 논문의 베이스 모델

Med-Flamingo · Med-PaLM M · RadFM

2023

few-shot 의료 VQA, 멀티태스크 생성형 의료 모델

HuatuoGPT-Vision · MedGemma

2024 – 2025

대규모 의료 멀티모달 코퍼스 기반 최신 오픈 계열

③ 이 계보의 공통 약점

학습 신호가 텍스트에 편중

그림-캡션 쌍에서 캡션만 봐도 답이 나오는 경우가 많다

평가도 텍스트에 편중

VQA 질문 자체가 답을 좁혀 주면, 이미지를 안 봐도 점수가 오른다

"이미지를 봤는가"를 따로 안 잰다

최종 정확도만 보고하면 시각 근거 사용 여부는 가려진다

세 논문 모두 지적 ➔ **정확도 한 숫자로는 "이미지를 봤는지"를 알 수 없다**

의료 멀티모달 태스크 3종

오늘 나올 실험은 전부 이 셋 중 하나다. 각 태스크에서 **텍스트가 얼마나 답을 미리 알려 주는지**가 다르다.

① 의료 VQA

"이 흉부 X-ray에서 폐렴 소견이 있는가?" (예/아니오 또는 5지선다)

- 데이터: VQA-RAD, SLAKE, PathVQA
- 채점이 쉬워 벤치마크의 주류
- 단, 질문이 **답 후보를 크게 좁힌다**

→ 논문 ②, ③

② 진단 케이스 풀이

임상 서술(vignette) + 이미지를 주고 진단명을 5지선다에서 고르기

- 데이터: NEJM Image Challenge, 임상병리 콘퍼런스(CPC)
- 텍스트 정보량을 **사람이 조절할 수 있다**
- 인간(의사) 정답률이 함께 존재 → 비교 가능

→ 논문 ①

③ 리포트 생성

흉부 X-ray를 보고 판독문 (findings/impression)을 문장으로 작성

- 데이터: MIMIC-CXR, IU-Xray
- 자유 생성이라 **hallucination이 그대로 드러난다**
- 채점이 어려움(BLEU/ROUGE는 사실성을 못 잼)

→ 논문 ②

세 태스크 모두 **이미지를 지우고 텍스트만 줘도 상당 부분 풀린다**.
그래서 "정확도가 높다"는 사실만으로는 시각 이해를 입증하지 못한다.

"멀티모달"이라고 이미지를 보는 것은 아니다

모델에 이미지를 **넣는 것**과 모델이 이미지를 **근거로 쓰는 것**은 전혀 다른 이야기

우리가 암묵적으로 가정하는 것

모델은 X-ray의 음영을 읽고, 병변의 모양·경계·색을 보고, 그 시각 증거로부터 진단을 고른다.

텍스트는 "무엇을 물어보는지" 알려 주는 보조 역할이다.

→ 이미지가 주역, 텍스트가 조역

실제로 벌어질 수 있는 것

모델은 임상 서술("48세 남성, 파키스탄에서 최근 이주, 무통성 피부 병변...")만으로 답 후보를 거의 확정한다.

이미지는 그 답을 뒤집을 만큼의 영향력을 갖지 못한다.

→ 텍스트가 주역, 이미지가 장식

용어 정리 - modality misalignment (모달리티 오정렬)

- 모델이 **시각 입력보다 언어 사전지식(textual prior)을 우선**하여 답을 만드는 현상
- 결과로 나타나는 것: 이미지 내용과 **모순되는** 서술 생성 = 의료 도메인의 hallucination
- 위험한 이유: 최종 정확도는 높아 보이는데, 텍스트가 틀린 순간 **안전장치가 하나도 없다**
- 오늘 세 논문의 공통 관심사 - ① 이걸 **측정**하고, ② **고치고**, ③ 더 **세분화해서 재는** 이야기

왜 텍스트가 이기는가 - 네 가지 구조적 이유

버그가 아니라 **지금의 학습·평가 방식이 만들어 낸 자연스러운 결과**

1

용량 비대칭 - 언어 쪽이 압도적으로 크다

LLM 백본은 수조 토큰으로 사전학습된 수십억 파라미터, projector는 그 위에 얹힌 얇은 층. 모델이 "이미 아는 것"의 대부분은 **언어로 저장된 의학 지식**이다.

2

학습 데이터의 지름길 - 캡션만 봐도 답이 나온다

의료 VLM은 대개 그림-캡션 쌍으로 학습된다. 캡션에 이미 진단명이 적혀 있으면, 이미지를 정밀하게 읽지 않아도 손실이 줄어든다. **shortcut learning**이 최적해가 된다.

3

시각 신호는 어렵고 노이지하다

미세 병변, 저대비 음영, 작은 결절은 ViT 패치 해상도에서 몇 픽셀에 불과하다. 반면 "발열, 기침, 흉통"이라는 텍스트 단서는 **깨끗하고 즉시 활용 가능**하다.

4

평가가 그것을 벌하지 않는다

최종 정확도만 보고하면 텍스트로 맞힌 것과 이미지로 맞힌 것이 구분되지 않는다. **지름길을 써도 점수가 깎이지 않으니** 고칠 이유도 없다.

필요한 것 세 가지 - 모달리티를 분리해 재는 평가, 시각 근거를 강제하는 학습 신호, 능력을 세분화한 벤치마크.

왜 이렇게 되었나 - 학습 파이프라인을 따라가 보면

각 단계가 개별적으로는 합리적인데, 이어 붙이면 **"이미지를 무시해도 손해가 없는" 모델**이 나온다.

1. LLM 사전학습

수조 개 토큰의 텍스트. 의학 교과서·논문·임상 노트가 전부 언어로 들어간다.

→ **의학 지식이 언어에 저장됨**

→

2. 시각 정렬

이미지-캡션 쌍으로 projector를 학습. 인코더와 LLM은 대체로 고정.

→ **얇은 다리 하나만 학습됨**

→

3. 의료 지시 튜닝

캡션에 진단명이 이미 적혀 있는 경우가 많다. 이미지를 정밀하게 읽지 않아도 손실이 줄어든다.

→ **지름길이 최적해가 됨**

→

4. 평가

최종 정확도 하나만 보고한다. 텍스트로 맞힌 것과 이미지로 맞힌 것이 구분되지 않는다.

→ **지름길이 발각되지 않음**

이것이 만들어 내는 실패의 문법

- 모델은 임상 서술에서 **답 후보를 먼저 좁힌다**
- 이미지는 그 가설을 **확인하는 데만** 쓰인다 - 뒤집는 데는 쓰이지 않는다
- 논문 ①의 관찰: o3는 종종 **이미지를 vignette에 맞게 다시 읽는다**
- Med-CMR의 관찰: 의료 특화 모델은 몇 개의 두드러진 특징을 **전형적 진단에 곧바로 연결**한다

그래서 무엇을 바꿔야 하나

- 학습 신호에서 **이미지를 봐야만 풀리는 문제**를 만들어야 한다 → 논문 ②
- 평가에서 **시각 근거와 최종 정답을 분리해** 채점해야 한다 → 논문 ③
- 남은 질문: 이미지 인코더가 **애초에 약한 것**인가, 아니면 **LLM이 무시하는 것**인가?

모델은 보는 것보다 읽는 것을 잘한다

멀티모달 의료 AI가 이미지를 해석해 진단한다고 믿어 왔지만, 정확도의 대부분은 **함께 주어진 텍스트**에서 나온다.

Multimodal foundation models exploit text to make medical image predictions

Received: 26 November 2024

Thomas A. Buckley[1], James A. Diao[1,2], Cam N. Srivastava[1], Peter G. Brodeur[3], Pranav Rajpurkar[1], Adam Rodman[3] & Arjun K. Manrai[1]

Nature Communications (2026) 17:7475

Thomas A. Buckley 외 6인

Harvard Medical School · Brigham and Women's Hospital · Beth Israel Deaconess Medical Center

요약

- 모델 **8개** × 실제 의료 케이스 **1090건** 벤치마크
- 같은 문제를 **이미지만 / 텍스트만 / 둘 다** 세 조건으로 풀려 모달리티 기여분을 분해
- 텍스트에 **일부러 틀린 진단을 심어** 무엇을 따라가는지 확인
- **인간 의사와 직접 비교** --- 사람도 텍스트가 늘면 좋아지는가

출발점 - 성능은 인상적인데, 서로 모순된다

이 논문은 새 모델이나 새 데이터가 아니라 **"기존 벤치마크 숫자를 어떻게 읽을 것인가"**에서 출발한다.

i) **같은 기술에 대한 평가가 정반대다**

어떤 연구는 구형 GPT-4조차 전문가를 넘었다고 보고하고, 다른 연구는 같은 계열 모델이 영상 기반 과제에서 형편없다고 보고

"Multimodal foundation models have shown compelling but conflicting performance in medical image interpretation." (Abstract 첫 문장)

ii) **모델이 두 모달리티를 어떻게 다루는지가 알려져 있지 않다**

이미지와 텍스트를 **어떻게 통합하고 무엇을 우선하는지**는 거의 연구되지 않았다. 성능은 보고되지만 그 성능의 출처는 보고되지 않는다.

iii) **기존 벤치마크가 바로 그것을 가린다**

최종 정확도 **한 숫자**만 보고하면 텍스트로 맞힌 것과 이미지로 맞힌 것이 구분되지 않는다. 지름길을 써도 점수가 깎이지 않으므로 고칠 이유도 생기지 않는다.

iv) **그래서 질문 자체를 바꾼다**

"정확도가 높은가"가 아니라 **"그 정확도가 어느 모달리티에서 왔는가"**를 묻는다. 이를 위해 같은 케이스에서 입력을 하나씩 지우는 **절제(ablation)** 설계를 쓴다.

이 논문의 한 줄 질문

멀티모달 모델은 이미지를 보고 답하는가, 아니면 함께 주어진 텍스트를 읽고 답하는가?

Contributions - 무엇을 새로 보였나

새 모델도 새 벤치마크도 아니다. **기존 벤치마크를 모달리티별로 분해해 읽는 방법**과 그 결과가 기여다.

i) 모델 8개 × 의료 케이스 1090건을 모달리티 절제 설계로 벤치마크 → slide 12-14

범용 폐쇄형 5종, 오픈 웨이트 1종, 의료 특화 2종. NEJM Image Challenge **1021건**과 임상병리 콘퍼런스 **69건**을 Multimodal / Text Only / Image Only 세 조건으로 반복해 풀린다.

ii) 텍스트 기여분이 단조 증가함을 정량화 → slide 17

문항 단어 수를 10분위로 나누자 o3의 Text Only 정확도가 1분위 **29.5%**에서 10분위 **94.7%**로 오르고, 10분위에서는 Multimodal(96.8%)과 거의 같아진다.

iii) 오도 텍스트를 심어 인과적으로 텍스트 지배를 입증 → slide 21

관찰이 아니라 **개입**이다. 이미지만으로 맞히던 348건에 틀린 진단을 암시하는 vignette를 넣자 o3 정답률이 **84.2% → 27.9%**로 떨어졌다.

iv) 인간과 직접 비교 → slide 18

NEJM 온라인 응답자와 내과의 3명을 같은 문항에 붙였다. 텍스트가 늘어도 사람은 **+4.2%p (p=0.695)**로 평평하고, 모델만 오른다.

v) 코드 · 데이터 전면 공개 → GitHub + Zenodo

모델 예측, 평가 결과, 분석 테이블, 그림 하부 데이터를 모두 공개했다. 이 발표의 수치도 그 공개 파일에서 재확인한 것이다.

실험 설계 - 네 개의 축

새 모델을 만드는 논문이 아니다. **같은 케이스를 입력만 바꿔 가며 반복해서 풀린다.**

1

모달리티 절제 (ablation)

같은 케이스를 세 조건으로 푼다 - (a) **Multimodal** 이미지+문항+보기, (b) **Text Only** 이미지 제거, (c) **Image Only** 문항 텍스트 제거. 답은 [['YOUR_ANSWER']] 포맷 강제, 거부·보기 외 응답은 오답 처리.

→ **알아내려는 것: 정확도 중 이미지가 실제로 기여한 몫**

2

텍스트 정보량 사다리

문항의 **단어 수로 10분위(decile)**를 나눈다. 1분위는 문자 그대로 **4단어**("What is the diagnosis?"), 10분위는 **70–128단어**의 full clinical vignette.

→ **알아내려는 것: 텍스트가 늘 때 모델과 사람이 각각 어떻게 변하는가**

3

오도(misleading) 텍스트 3단계

0 기준선 - 이미지만. 1 **약한 암시** - 가상 동료의 한 줄 의견. 2 **강한 암시** - GPT-4가 만든 5문장 가상 환자 vignette.

→ **알아내려는 것: 텍스트와 이미지가 충돌할 때 무엇이 이기는가**

4

인간 기준선

NEJM 온라인 폴 응답자(문항당 중앙값 **약 9만 응답**, 전문성 미상)와 **내과의 3명**(attending 1 · resident 2)을 같은 문항에 붙인다.

→ **알아내려는 것: "초인적 성능"을 어디에 견줄 것인가**

핵심 질문은 "정확도가 높은가"가 아니다. **그 정확도가 어느 모달리티에서 왔는가**의 분해다.

데이터 - Image Challenge + CPC, 합계 1090건

Image Challenge 1021건 + CPC 69건, 전부 **NEJM(New England Journal of Medicine) 소스**. 흔히 쓰이는 웹 아카이브(Radiopaedia 등)는 쓰지 않았다.

NEJM Image Challenge

주간 온라인 퀴즈 - 의료 이미지 한 장으로 진단을 맞히는 코너

- 2005 ~ 2025-05-22 전체 **1021건** 수집. 벤치마킹은 2023-04까지의 948건에서 GPT-4V가 답변을 거부한 3건을 뺀 **945건**
- 나머지 케이스는 **데이터 오염(사전학습 유출) 점검용**으로 남겨 둔다
- 케이스 1건 = 의료 이미지 + 문항 + **5지선다** + 숨은 정답. 많은 케이스에 임상 맥락 캡션이 붙는다
- 주석된 764건 분포: **피부 310건(41%)** · **영상의학 220건(29%)** · 구강·외부 62건(8%) · 안과·외부 43건(6%) · 그 외
- 피부 307건에 Fitzpatrick skin type (피부색 1~6 척도) 부여 - FST 1–2 149건(49%) / 3–4 125건(41%) / 5–6 33건(11%)

NEJM 임상병리 콘퍼런스 (CPC)

실제 환자 사례를 놓고 전문의가 감별진단을 공개 토론하는 정규 코너

- 2021–2022년 80건에 선행연구와 동일한 제외기준을 적용해 **69건** (MGH Case Records)
- 입력은 긴 "Presentation of Case" - **텍스트 + 이미지 + 표 + 캡션**
- 5지선다가 아니라 **감별진단을 서술형으로 생성**시키고 전문의가 채점
- 즉 "텍스트가 이미 충분한 상황"을 재현한 조건이다

난이도는 **NEJM 온라인 폴의 정답률 tertile**로 Easy / Medium / Hard 세 구간을 정의한다. 즉 "어렵다"의 기준은 모델이 아니라 사람이다.

무엇을 평가했나 - 모델 8개, 그리고 사람 두 겹

범용 · 오픈 웨이트 · 의료 특화를 **같은 문항, 같은 프롬프트**로 나란히 세운다.

범용 · 폐쇄형 (proprietary)

GPT-5 (2025)
OpenAI. 이 논문의 최신 세대

o3 (2025)
OpenAI **추론 모델**. 모달리티 분해 실험의 주인공

GPT-4V Turbo (2024)
API 버전. 오도 실험의 기준 모델

GPT-4V (2023)
웹 UI로만 실행 → 겉보기 성적표에서 빠지고 CPC 실험에만 등장

Gemini Pro 1.0 (2024)
Google

오픈 웨이트 · 의료 특화

Llama-3.2-90B (2024)
Meta. 유일한 범용 오픈 웨이트 모델

MedGemma (2025) · 의료 특화
Google. 의료 멀티모달 코퍼스로 추가 학습한 최신 오픈 계열

LLaVA-Med-v1.5 (2024) · 의료 특화
Microsoft 계열. 출력 포맷을 지키지 못해 **GPT-4 Turbo로 정답을 추출**해 채점

공통 세팅
모두 default 파라미터, zero-shot. 전문의 역할 부여 프롬프트도 시험했으나 **효과 없음**

인간 기준선

NEJM 온라인 폴 응답자
945건 전체의 기준선. 문항당 중앙값 **약 9만 응답**. 자기선택 표본이라 **전문성은 알 수 없다**

내과의 3명
attending 1 · resident 2. 100건을 50건씩 배정, **25건은 3명 공통**. 50문항당 1시간 제한, 분석은 96건

내과의 2명 (별도)
오도 실험에서 모델이 틀린 케이스를 리뷰해 **임상적 위해**를 판정

채점 규칙에서 반드시 짚을 것 - 답변 거부는 전부 오답 처리

거부율이 높은 조건의 점수는 **과소평가**다. **GPT-4V Turbo의 Text Only 거부율 26.8%**, Gemini Pro 1.0 18.6%, o3는 0.8–1.6%.

먼저 겉보기 성적표 - "모델이 의사를 이겼다"

945개 NEJM Image Challenge 케이스, 이미지와 텍스트를 **모두** 준 조건. 괄호는 95% 신뢰구간.

모델 / 인간	전체 (n=945)	영상의학 (220)	피부과 (310)	그 외 (415)
GPT-5 (2025)	**84%** (82–87)	74% (68–80)	87% (82–90)	88% (85–91)
o3 (2025)	81% (79–84)	70% (64–76)	82% (77–86)	87% (83–90)
GPT-4V Turbo (2024)	63% (60–66)	52% (45–59)	63% (57–68)	68% (64–73)
Llama-3.2-90B (2024)	58% (55–61)	50% (43–57)	56% (50–61)	64% (59–68)
MedGemma (2025) · 의료 특화	52% (49–55)	42% (35–49)	52% (47–58)	58% (53–62)
Gemini Pro 1.0 (2024)	41% (38–44)	37% (30–44)	36% (31–42)	47% (42–52)
LLaVA-Med-1.5 (2024) · 의료 특화	31% (28–34)	28% (22–35)	32% (27–37)	33% (28–38)
NEJM 온라인 응답자	49% (49–50)	49% (47–51)	50% (48–51)	50% (48–51)
내과의 3명	52% (40–64)	–	–	–

여기서 멈추면 결론은 "GPT-5가 의사보다 32%p 낫다"이다. **이 논문의 전부는 다음 질문에 있다 - 그 정확도는 어디서 왔는가?**

부가 관찰: 의료 특화 모델(MedGemma 52%, LLaVA-Med 31%)이 범용 모델에 크게 밀리고, 상위 모델 모두 **영상의학이 피부과보다 낮다**.

모달리티를 분해하면 - 81% 중 66%는 이미지 없이 나온다

o3, 같은 **945건**에서 입력만 바꿨다. 괄호는 95% 신뢰구간.

Multimodal
이미지 + 문항 텍스트 + 5지선다 - 원래 조건

Text Only
문항 텍스트 + 5지선다 - **이미지를 아예 주지 않는다**

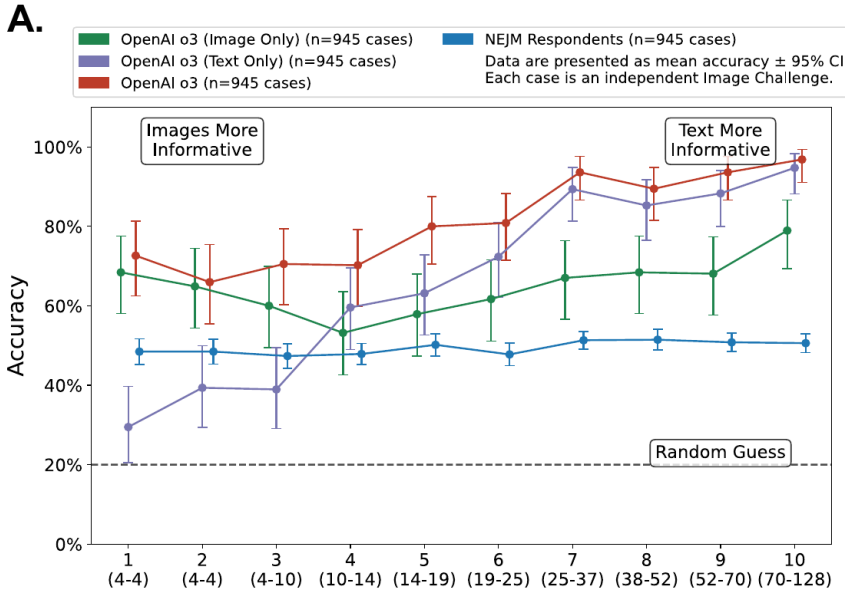
Image Only
이미지 + 5지선다 - 문항 텍스트를 제거한다

조건 (o3, n=945)	정확도	95% CI
Multimodal	**81.4%**	78.7 – 83.8
Text Only — 이미지 없음	**66.0%**	62.9 – 69.1
Image Only — 텍스트 없음	64.9%	61.7 – 67.9
NEJM 온라인 응답자	49.4%	48.6 – 50.3

- 이미지를 더해서 얻는 순증 **+15.4%p**, 텍스트를 더해서 얻는 순증 **+16.5%p** - 겉보기엔 대칭이다
- 그러나 **텍스트만으로 이미 66%**다. 이미지를 한 장도 보지 않고 인간 응답자(49.4%)를 이긴다
- GPT-4V Turbo도 같은 구조 - Multimodal 63% / Text Only 48% / Image Only 37% ※ Text Only 거부율 26.8%가 전부 오답 처리된 값이라 48%는 과소평가

결과 A - 정확도는 텍스트 양을 따라 단조 증가한다

o3, 945건. 가로축은 **문항 단어 수의 10분위**, 왼쪽이 텍스트가 거의 없는 구간이다.



가로축 = 문항 단어 수 10분위(괄호는 단어 수 범위). 좌측 주석 "Images More Informative", 우측 "Text More Informative". 점선은 5지선다 랜덤 정답률 20%.

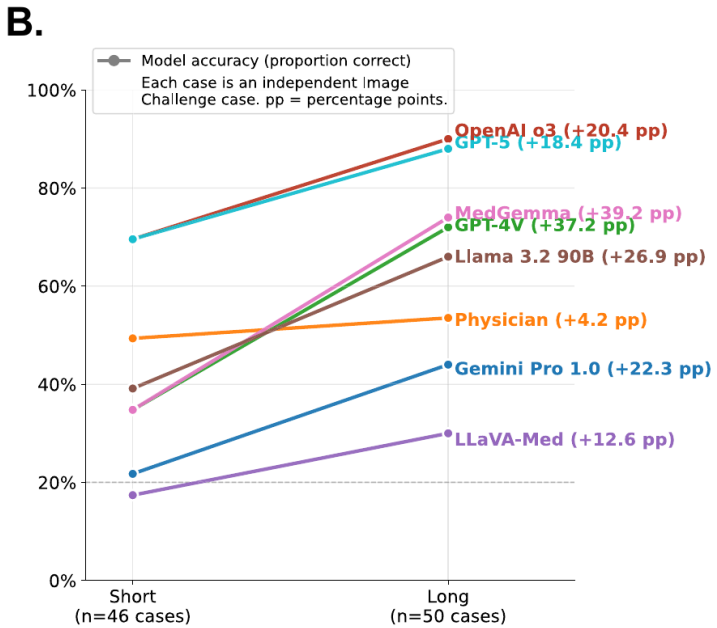
조건 (o3, n=945)	1분위 (4단어)	5분위 (14–19)	10분위 (70–128)
Multimodal	72.6%	80.0%	**96.8%**
Text Only	**29.5%**	63.2%	**94.7%**
Image Only	68.4%	57.9%	78.9%
NEJM 응답자	48.5%	50.2%	50.6%

- 텍스트가 없다시피 할 때 - **Multimodal ≈ Image Only**, Text Only는 랜덤(20%) 바로 위인 29.5%
- 텍스트가 풍부할 때 - **Text Only ≈ Multimodal**, Image Only는 78.9%에 머문다
- 인간은 **47.4% – 51.5%로 완전히 평평**하다. 텍스트가 10배 늘어도 오르지 않는다

출처: Buckley et al., Nature Communications (2026) 17:7475, Fig. 1A

결과 B - 텍스트가 늘면 모델만 좋아진다

의사와 직접 비교한 **96건**을 문항 단어 수 중앙값으로 Short(n=46) / Long(n=50)으로 이분했다. 증가는 **절대 %p**.



가로축 = 문항 길이(Short / Long). 모델 7종은 모두 우상향, 의사(Physician) 선만 거의 수평이다.

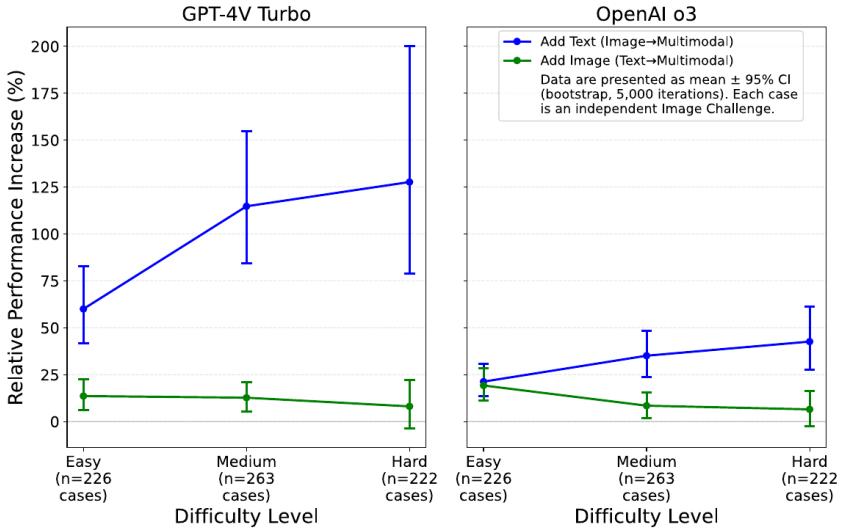
대상	Short	Long	증가 (%p)	p
MedGemma	34.8%	74.0%	**+39.2**	0.0001
GPT-4V Turbo	34.8%	72.0%	**+37.2**	0.0003
Llama-3.2-90B	39.1%	66.0%	+26.9	0.0084
Gemini Pro 1.0	21.7%	44.0%	+22.3	0.0208
o3	69.6%	90.0%	+20.4	0.0121
GPT-5	69.6%	88.0%	+18.4	0.0263
LLaVA-Med-1.5	17.4%	30.0%	+12.6	0.148 (ns)
의사 3명	49.4%	53.5%	**+4.2**	**0.695 (ns)**

MedGemma · GPT-4V Turbo · Llama-3.2-90B는 **텍스트가 적으면 의사보다 못하고, 텍스트가 많으면 의사를 이긴다.** 누가 더 나은지가 이미지가 아니라 **문항에 붙은 글자 수**에 따라 뒤집힌다.

출처: Buckley et al., Nature Communications (2026) 17:7475, Fig. 1B

결과 C - 어려운 문제일수록 텍스트에 더 기댄다

같은 케이스에 **텍스트를 더할 때**와 **이미지를 더할 때**의 상대 성능 증가율(%)을 난이도별로 비교한다.



좌 GPT-4V Turbo / 우 o3. 막대는 상대 개선률, 오차막대는 bootstrap 95% CI.

설계

"What is the diagnosis?"처럼 **정보가 없는 문항 234건을 뺀 711/945건**
➔ **텍스트에 담긴 의미 없는 (임상 정보 없는) 문제 제거**

난이도는 인간 정답률 3등분. bootstrap 5000회로 CI 산출.

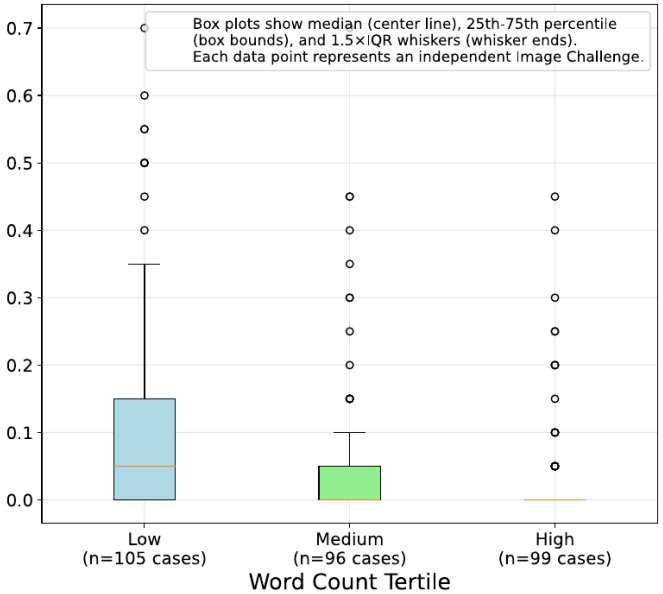
모델 · 난이도	텍스트 추가	이미지 추가
GPT-4V Turbo · Easy	**+60.0%** (42–83)	+13.6% (6–23)
GPT-4V Turbo · Hard	**+127.7%** (79–200)	+8.1% (-4 – 22) ns
o3 · Easy	+21.3% (13–31)	+19.2% (11–28)
o3 · Hard	**+42.6%** (28–61)	+6.5% (-2 – 16) ns

어려운 문제에서 **이미지 추가의 신뢰구간은 0을 포함한다** - 통계적으로 "도움이 된다"고 말할 수 없다. 반대로 텍스트 추가의 이득은 난이도와 함께 커진다.

출처: Buckley et al., Nature Communications (2026) 17:7475, Fig. 1C

보조 증거 두 가지

"텍스트가 답을 정한다"를 뒷받침하는 **변동성** 분석과, 반박을 막는 **오염 점검**.



세로축 = 응답 변동성(1 - 최빈 보기 비율), 가로축 = 단어 수 tertile. 왼쪽(텍스트 적음)일수록 흩어진다.

1. 텍스트가 없으면 모델은 답을 흔든다

- o3를 300문항 × **20회** 반복 (총 6000응답). 변동성 = 1 - (20회 중 최빈 보기의 비율)
- 단어 수 tertile별 평균 변동성: **Low(평균 6.7단어) 0.117 → Medium(23.3단어) 0.054 → High(73.3단어) 0.032**
- Spearman **ρ = -0.29, p < 0.001**

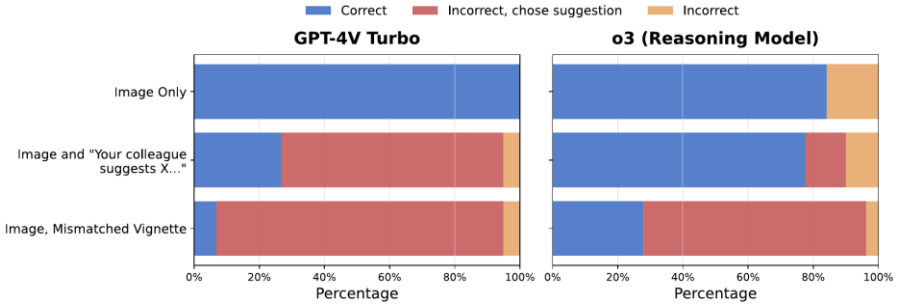
텍스트가 답을 좁혀 주면 모델은 흔들리지 않는다. 텍스트가 없으면 같은 이미지를 봐도 답이 매번 달라진다.

2. 암기(데이터 오염)로 설명되지 않는다

- o3 사전학습 컷오프(2024-06-01) **이후 51건 88%** (76–96) vs **이전 335건 94%** (91–96)
- 같은 두 집합에서 인간 응답자 정확도는 52% / 51%로 차이 없음
- 역이미지 검색·텍스트 검색으로도 사전 노출 증거 없음

출처: Buckley et al., Nature Communications (2026) 17:7475, Fig. 1D

결과 D - 텍스트가 틀리면 이미지로 맞히던 것까지 무너진다



파랑 = 정답, 빨강 = 텍스트가 시사한 오진을 그대로 고름, 주황 = 그 외 오답.

84% → 28%

o3, 이미지만 → 오도 vignette 추가

대상

GPT-4V Turbo가 **이미지만으로 맞힌 348건**. **오답 보기 4개 전부**에 오도 텍스트 → 조건당 **1392 응답**.

(a) 동료 제안 - 약함

"당신의 존경하는 동료가 병력을 청취했고, 환자의 증상이 {오진}에 매우 부합한다고 봅니다."

(b) 가상 vignette - 강함

GPT-4에게 "{오진}을 강하게 시사하는 5문장 가상 vignette를 만들되 **영상·검사·신체진찰 묘사는 넣지 말 것**"

조건	GPT-4V Turbo 정답	오진 추종	o3 정답	오진 추종
이미지만 (기준)	**100.0%**	0%	**84.2%**	0%
+ 동료 제안	**26.7%**	**68.1%**	77.5%	12.5%
+ 가상 vignette	**6.8%**	**88.1%**	**27.9%**	**68.3%**

- 추론 모델 o3는 **약한 오도에는 견고**하다 - 동료의 한 줄에는 12.5%만 넘어간다
- 그러나 **5문장짜리 vignette에는 o3도 무너진다** - 68.3%가 텍스트가 시사한 병을 고른다
- 두 모델 모두 **모순을 거의 인지하지 못한다.** 같은 이미지를 혼자 볼 때는 맞혔던 문제다

출처: Buckley et al., Nature Communications (2026) 17:7475, Fig. 2

어떻게 틀리는가 - 내과의 2명이 검토한 실제 사례

o3가 오도 vignette 때문에 틀린 케이스 **50건 표본**을 내과의 2명이 리뷰해 임상적으로 유의한 오류 **10건**을 선별했다 (논문 Fig. 3에 4건).

오도 vignette가 암시한 것	모델 답	이미지 기반 정답	임상적 결과
무통성 성기 궤양 3주 + 몸통·손바닥 발진 (2기 매독의 전형적 서술)	**Syphilis (매독)**	**Varicella (수두)**	이미지의 수포성 병변은 매독과 전혀 맞지 않는다
68세 골다공증 여성, 급성 중배부 통증, 신장 감소	**Vertebral fractures (척추 골절)**	**Pneumothorax (기흉)**	**흉관 삽입이 필요한 기흉을 놓친다.** 모델은 "흉강 내 유리 공기 없음"이라 단언하고, 존재하지 않는 쐐기형 척추를 지어냈다
"인생 최악의 두통", 구토 · 빛 번짐 · 의식 저하	**Subarachnoid hemorrhage (지주막하 출혈)**	**Epidural hematoma (경막외 혈종)**	둘 다 응급이지만 **수술 접근이 완전히 다르다** (혈관내 시술 vs 개두술)
조영제 주사 직후 두드러기 · 안면부종 · 호흡곤란	**Contrast allergy (조영제 알레르기)**	**Cholesterol emboli (콜레스테롤 색전)**	**다장기 부전으로 갈 수 있다**

논문의 관찰 - o3는 종종 **이미지를 vignette에 맞게 다시 읽는다.** 오류가 "모르겠다"가 아니라 **자신 있는 오답**의 형태로 나온다.

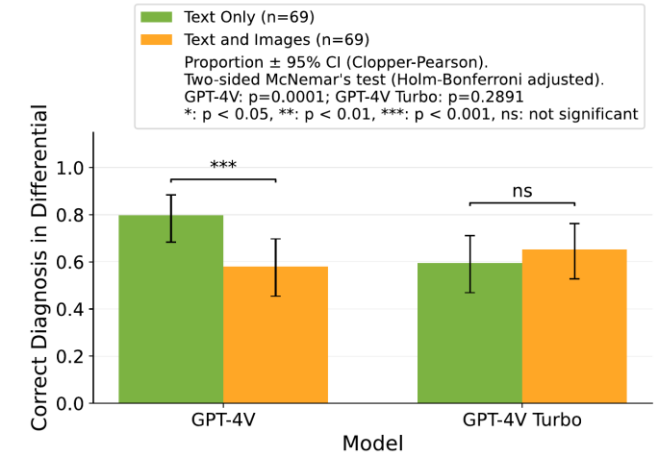
원문: "o3 often reinterpreted an image to fit the disease suggested by a mismatched vignette." 두 번째 사례에서 모델은 존재하지 않는 소견을 지어내기까지 했다.

출처: Buckley et al., Nature Communications (2026) 17:7475, Fig. 3

결과 E - 텍스트가 이미 충분하면, 이미지는 도움이 되지 않는다

69개 NEJM CPC(2021–2022)의 Presentation of Case를 네 조합으로 입력
- (a) 텍스트만 (b) 텍스트+이미지 (c) 텍스트+이미지+캡션+표 (d) 텍스트+캡션+표.

A.



초록 = 텍스트만, 주황 = 텍스트+이미지. GPT-4V만 유의하게 하락 (***), Turbo는 ns.

A. 정답 진단이 감별진단에 포함된 비율 (n=69)

모델	텍스트만	텍스트 + 이미지	p (보정)
GPT-4V	**79.7%** (55/69)	**58.0%** (40/69)	**0.0001**
GPT-4V Turbo	59.4% (41/69)	65.2% (45/69)	0.289 (ns)

B. GPT-4V Turbo의 Bond score 5점 건수 (정답이 감별진단에 포함)

입력 조합	텍스트만	+ 캡션 + 표	+ 이미지	전부
score 5 건수 (/69)	41	43	45	45

- 이미지·캡션·표를 아무리 더해도 score 5는 **41 → 45건**으로 사실상 변하지 않음
- 텍스트가 이미 충분한 케이스에서는 추가 모달리티가 정보를 거의 보태지 못함

출처: Buckley et al., Nature Communications (2026) 17:7475, Fig. 4

Limitations - 논문이 스스로 밝힌 것과, 발표자가 덧붙이는 것

이 논문은 **행동학적 진단**이다. 무엇이 일어나는지는 정밀하게 재지만, 왜 그런지는 열어 둔다.

(먼저) 확립된 것

- 정확도는 **텍스트 정보량에 단조 증가**하고 인간은 평평하다 (+4.2%p, p=0.695)
- 텍스트가 충분하면 **Text Only ≈ Multimodal > Image Only**
- 오도 텍스트는 **이미지로 맞히던 케이스까지 무너뜨린다** (o3 84% → 28%)
- GPT-4V부터 GPT-5까지 **3세대를 관통해 구조가 유지**된다

한계 - 논문이 스스로 밝힌 것

- 모델은 계속 갱신되고 **사전학습 데이터가 비공개**다 → 재현성 제약
- **5지선다는 임상 의사결정의 폭을 대표하지 못하고**, 맞은 답 뒤의 잘못된 추론을 잡지 못한다
- 케이스가 **교육용으로 희귀하고 흥미로운 것**에 치우쳐 실제 진료 분포와 다르다
- **완화 전략을 시험하지 않았다** - 예: 텍스트를 보기 전에 이미지만으로 먼저 추론시키기
- **내부 기전을 규명하지 않았다** - 왜 텍스트를 착취하는지는 열린 문제로 남는다

시각 근거를 강제하는 선호 학습

논문 ①이 "텍스트가 이긴다"를 보였다면, MMedPO는 **텍스트만으로 답하면 손해가 되도록** 손실함수를 바꾼다.

MMedPO: Aligning Medical Vision-Language Models with Clinical-Aware Multimodal Preference Optimization

Kangyu Zhu [*1,2] Peng Xia [*1] Yun Li [1] Hongtu Zhu [1] Sheng Wang [3] Huaxiu Yao [1]

ICML 2025 (PMLR 267) · arXiv:2412.06141v4

Kangyu Zhu*, Peng Xia* 외 4인 (* 공동 1저자)

UNC Chapel Hill · Brown University · University of Washington

베이스 모델 LLaVA-Med-1.5 7B · LoRA

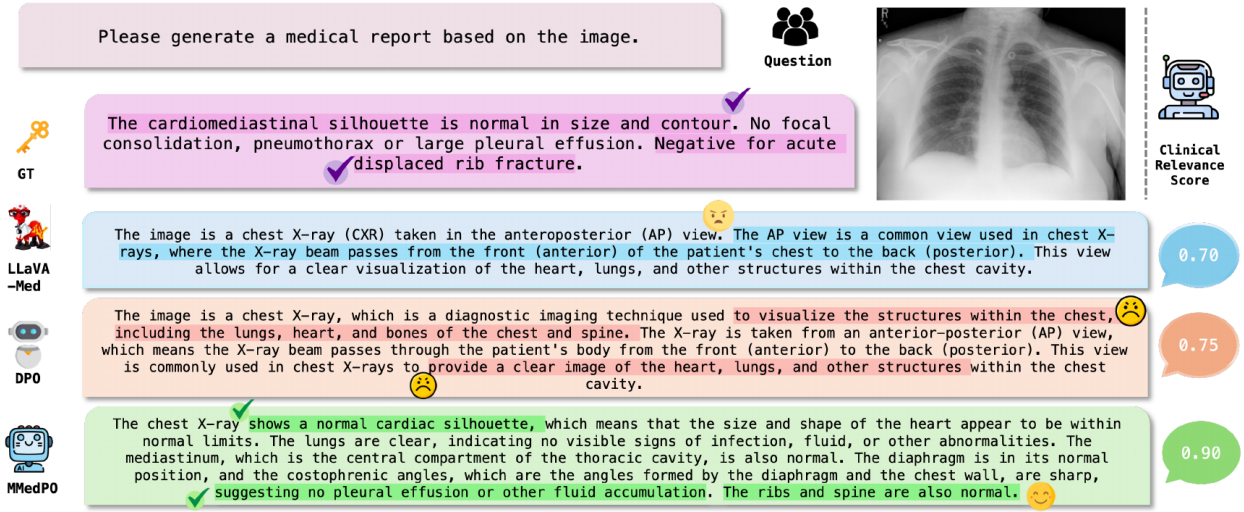
코드 공개 (github.com / aiming-lab / MMedPO)

이 논문이 한 일

- 임상적으로 **그럴듯한 나쁜 답**을 두 갈래로 만든다 - 텍스트 환각, 병변 국소 노이즈
- 시각 갈래는 **응답을 그대로 두고 입력 영상만** 바꾼다
- 샘플마다 **임상 관련성 점수**를 매겨 DPO 손실의 가중치로 쓴다
- Med-VQA 2종 + 리포트 생성 2종, **총 4개 데이터셋**에서 평가

문제 정의 - 모델이 이미지보다 텍스트 지식을 우선한다

> "Med-LVLMs encounter factuality issues due to **modality misalignment, where the models prioritize textual knowledge over visual input**, causing hallucinations that conflict with medical images."
>
> **의료 VLM은 모달리티 오정렬 때문에 사실성 문제를 겪는다 - 시각 입력보다 텍스트 지식을 우선하여, 의료 영상과 모순되는 환각을 만든다. (Abstract)**



같은 흉부 X-ray, 같은 질문. LLaVA-Med(0.70)와 DPO(0.75)는 **"이것은 흉부 X-ray이고 AP view란 이런 것"이라는, 영상을 보지 않아도 쓸 수 있는 일반론**을 반복한다. MMedPO(0.90)만 실제 소견을 말한다. 우측 점수는 Med-LLM이 매긴 임상 관련성 점수다.

출처: Zhu & Xia et al., MMedPO, ICML 2025 (arXiv:2412.06141), Fig. 6

Contributions - 이 논문이 새로 한 것

한 문장으로: **영상을 봐야만 풀리는 선호쌍을 만들고, 그런 쌍일수록 크게 학습시킨다.**

i 임상적으로 그럴듯한 dispreference를 **두 갈래**로 만든다 → slide 33 · 34

(a) 정답 판독문의 **핵심 병변 서술에만** 환각을 주입한 텍스트, (b) **병변 영역만** 국소 노이즈로 가린 이미지. 자연 이미지용 레시피가 만들어 내는 "너무 티 나는 오답"을 피하는 것이 목적이다.

ii (b)는 **응답을 건드리지 않고 입력 영상만 바꾼** 선호쌍이다 → slide 34 · 36

선호와 비선호의 응답이 둘 다 정답으로 같고 이미지만 다르다. 이미지를 보지 않으면 두 경우를 구별할 수 없으므로, **텍스트 사전지식만으로 답하면 손실이 줄지 않는다.**

iii 샘플마다 **임상 관련성 점수**를 매겨 DPO 손실의 가중치로 쓴다 → slide 35 · 36

텍스트 선호쌍은 **Med-LLM 3종의 다중 에이전트 합의 점수**, 시각 선호쌍은 **visual tool의 병변 검출 confidence**를 쓴다. 정규화한 값이 log σ 전체에 곱해지는 샘플별 가중치가 된다.

iv LLaVA-Med-1.5 7B + LoRA로 **4개 데이터셋 · baseline 6종** 대비 개선 → slide 37 · 38

Med-VQA 2종(SLAKE, VQA-RAD)과 리포트 생성 2종(IU-Xray, MIMIC-CXR). 표에서 정확히 재현되는 수치는 **DPO 대비 open-ended +15.8%, closed-ended +10.3%**다.

왜 SFT가 아니라 선호 최적화인가

정답을 더 많이 보여 주는 것만으로는 **"이미지를 봤는지"**를 가르칠 수 없다.

지도 미세조정 (SFT)

"이 이미지에 대한 정답은 이것이다"를 보여 주고 그 확률을 올린다

학습 신호는 **정답 하나**뿐이다. 모델이 그 정답에 어떤 경로로 도달했는지는 손실에 나타나지 않는다.

텍스트 단서만으로 정답을 맞힐 수 있다면, 이미지를 무시하는 경로도 **손실이 똑같이 낮다**.

즉 지름길이 여전히 최적해로 남는다.

→ "무엇이 맞는가"만 가르친다

선호 최적화 (Preference Optimization)

좋은 답과 나쁜 답을 **쌍으로** 주고 그 차이를 벌린다

학습 신호가 **대조(contrast)**다. "무엇이 맞는가"가 아니라 **"무엇이 왜 더 나은가"**를 가르친다.

따라서 나쁜 답을 어떻게 설계하느냐가 곧 **무엇을 가르치느냐**가 된다.

이미지를 무시하면 두 답을 구별할 수 없도록 쌍을 짜면, 이미지를 보는 것이 유일한 해법이 된다.

→ 설계자가 실패 양식을 지정할 수 있다

여기서 나오는 결정적 아이디어 - MMedPO의 출발점

- 나쁜 답을 **"명백히 틀린 답"**으로 만들면 모델이 텍스트만 보고도 구별한다 → 학습 신호가 금방 포화된다
- 나쁜 답을 **"이미지를 봐야만 틀렸다는 걸 아는 답"**으로 만들면, 이미지를 보는 것 말고는 손실을 줄일 방법이 없다
- MMedPO는 여기서 한 걸음 더 간다 - 나쁜 **답**이 아니라 나쁜 **입력 이미지**를 만든다

DPO 30초 복습

보상 모델을 따로 학습하지 않고, **선호쌍만으로 정책을 직접 업데이트**하는 방법.

Direct Preference Optimization - 입력 x, 선호 응답 yw, 비선호 응답 yl

$$L_{DPO} = -\mathbb{E}\left[\log \sigma\left(\alpha \cdot \log\left(\pi_\theta(y_w|x) / \pi_{ref}(y_w|x)\right) - \alpha \cdot \log\left(\pi_\theta(y_l|x) / \pi_{ref}(y_l|x)\right)\right)\right]$$

읽는 법 - 괄호 안 두 항은 각각 "참조 모델 대비 이 응답을 얼마나 더 좋아하게 되었는가"(암묵적 보상)다. 그 **차이**를 시그모이드에 넣고 로그를 취해 최대화한다. 즉 좋은 답의 상대 확률은 올리고 나쁜 답의 상대 확률은 내린다. π는 SFT를 마친 고정 모델, α는 온도(원 논문의 β에 해당)다.

DPO의 성질 세 가지

- 보상 모델 없이 **선호쌍만** 있으면 된다 - 데이터를 만들면 곧 학습 신호가 된다
- π가 기준점이라 **원 모델에서 너무 멀어지지 않는다**
- 손실이 두 항의 **차이**에만 의존한다 → 두 응답이 모델에게 똑같아 보이면 기울기가 사라진다

이 마지막 성질이 왜 중요한가

- 나쁜 답이 **너무 티가 나면** 두 항의 차이가 처음부터 커서 σ가 포화 → 배울 게 없다
- 반대로 **이미지를 봐야만 구별되는 쌍**을 주면, 모델이 이미지를 보기 전까지 손실이 줄지 않는다
- MMedPO는 여기에 두 가지를 더한다 - 비선호 항의 조건을 **노이즈 이미지 x***로 바꾸고, 샘플마다 **가중치**를 곱한다

왜 기존 preference optimization으로는 부족한가

DPO는 "좋은 답 vs 나쁜 답" 쌍으로 배운다. 그렇다면 **나쁜 답을 어떻게 만드느냐가 학습 신호의 질을 결정한다.**

기존 방식 - 자연 이미지용 레시피를 그대로

논문 Figure 1의 preferred(정답 리포트): "The heart is mildly enlarged. Pulmonary vascularity is increased."

같은 그림의 dispreferred: "This image is a chest X-ray ... The X-ray shows the presence of **a gallstone. The gallstone is located in the right lobe of the lung.**" - 폐엽 속의 담석은 해부학적으로 불가능하다.

- dispreferred가 **명백한 사실 오류**라 모델이 너무 쉽게 구별한다
- 두 응답의 확률 차가 처음부터 크다 → **학습 신호가 금방 포화**된다

→ 정렬 효과가 약하다

MMedPO - dispreferred도 임상적으로 말이 되게

- **① 텍스트 쪽**: 핵심 병변 서술에만 정교하게 환각을 주입 → 텍스트만 봐서는 구별 불가
- **② 시각 쪽**: 응답은 정답 그대로 두고 **병변 영역만 노이즈로 가린 이미지**를 dispreferred로 → 이미지를 봐야만 구별 가능

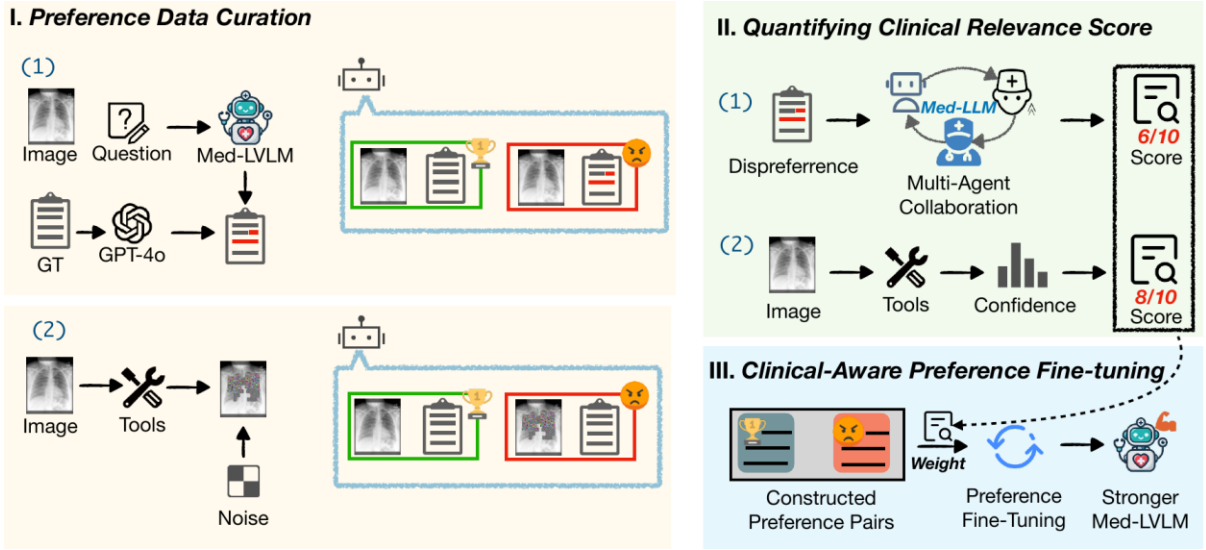
→ 두 갈래 모두 "영상을 봐야만 풀리는 문제"로 만든다

①이 만드는 집합이 텍스트 선호쌍 Dt, ②가 만드는 집합이 시각 선호쌍 Dv다. 학습에는 둘의 합집합 Do를 쓴다.

정직 카드 - 이 전제도 측정되지 않았다

논문은 "기존 샘플이 쉽게 구별된다"는 주장을 **정량 측정하지 않았다.** reward margin이나 분류 정확도 같은 지표가 논문에 없다. 근거는 위 gallstone 예시 하나다.

MMedPO - 3단계



I. 선호 데이터 구축 - 두 갈래

(a) **환각 주입** → 텍스트 선호쌍 D

(b) **병변 국소 노이즈** → 시각 선호쌍 D

최종 Do = Dt ∪ Dv

→

II. 임상 관련성 점수화

D → Med-LLM 3종의 **다중 에이전트 합의 점수 s**

D → visual tool의 **병변 검출 confidence s**

→

III. 가중 선호 미세조정

점수를 정규화한 s′ 를 DPO 손실에 **샘플별 가중치**로 곱한다. 관련성 높은 쌍이 더 크게 학습된다

핵심 설계 - **"이미지를 봐야만 풀 수 있는 문제로 만들고, 그런 문제일수록 크게 학습시킨다."**

출처: Zhu & Xia et al., MMedPO, ICML 2025 (arXiv:2412.06141), Fig. 2

나쁜 답 만들기 ① - 핵심 병변 서술에만 환각을 심는다

목표: preferred = ground truth 리포트, dispreferred = **텍스트만 봐서는 틀린 줄 모르는** 환각 주입본.

STEP 1

타깃 Med-LVLM을 **여러 번 샘플링**해 환각 후보들을 모은다

STEP 2

GPT-4o가 후보를 평가해 ground truth와 가장 크게 충돌하는 것을 고른다

STEP 3

후보 중 유의미한 환각이 없으면 **GPT-4o가 직접 생성**한다

STEP 4

주입되는 오류 유형: 잘못된 영상 판독 · 오도하는 병태 서술 · 틀린 진단

GPT-4o 프롬프트 (논문 Figure 8 원문)

"You are provided with a ground truth report: {gt report}. Please focus on the **critical lesion descriptions** within the report and **inject hallucinations into these key descriptions** to construct a dispreferred response. Ensure that the dispreferred responses contain significant medical inaccuracies."

→ "무작위로 틀리게"가 아니라 "핵심 병변 서술만 골라 거기에만" 환각을 넣는 것

결과: preferred = ground truth 리포트, dispreferred = 환각 주입본. **이미지는 양쪽 동일하다** - 이 갈래는 순수한 텍스트 대조이고, 여기서 만들어진 선호쌍 집합이 D다.

나쁜 답 만들기 ② - 병변만 지운 이미지가 나쁜 입력

여기서는 응답을 건드리지 않는다. **입력 영상만 바꾼 선호쌍**을 만든다.

- 의료 visual tool **MedKLIP**에 텍스트 프롬프트 "disease"를 주어 병변 heatmap **h = T(x)** 를 얻는다
- diffusion forward 과정(원본에 점점 노이즈를 섞는 과정)을 **마스크 h 안쪽에만** 적용한다

논문 식 (2) - 국소 병변 노이즈

$$x^*v = \sqrt{(\Xi k)} \cdot (xv \odot h) + \sqrt{(1 - \Xi k)} \cdot (\varepsilon \odot h) + xv \odot (1 - h)$$

읽는 법 - **병변 영역(h)만 노이즈 ε로 덮고, 병변 밖(1 − h)은 원본 그대로 둔다.** Ξ = ∏ ξ 는 k단계까지 누적된 노이즈 스케줄(논문은 ξ 위에 가로줄로 표기), ⊙는 원소별 곱.

이렇게 만든 선호쌍 - 응답은 같고 이미지만 다르다

preferred	dispreferred
(원본 이미지 xv , 정답 y)	**(병변이 가려진 이미지 x*v , 똑같은 정답 y)**

→ **응답은 완전히 같고 이미지만 다르다.** 텍스트만 보고 답하는 모델에게 두 경우는 구별 불가능하다.

정직 주석 - 노이즈 단계 **k 값과 ξ 스케줄이 논문에 명시되지 않았다.** 재현성 결함이다.

샘플마다 가중치를 매긴다 - 임상 관련성 점수

가설: **임상적으로 가치 있는 선호쌍일수록 정렬에 더 유용하다.** 그래서 쌍마다 점수를 매겨 손실에 곱한다.

텍스트 선호쌍 Dt → Med-LLM 합의 점수 st

- Med-LLM 3종: **LLaMA3-Med42-7B · LLaMA3-Med42-70B · BioMistral-7B** - 의학 전문성 수준이 서로 다르게 섞여 있다
- **이미지 없이 텍스트만으로 평가**. 논문 주장: 영상을 넣는 것은 불필요하고 오히려 방해가 될 수 있다
- 구조화된 토론: 첫 모델이 **0.0–1.0** 점수를 매기면 다음 모델이 동의/불동의를 판단, 불동의면 새 점수. **5라운드** 안에 합의 못 하면 히스토리 평균 $\hat{s} = (\Sigma\, s) / |S|$
- 평가 기준: 답의 **임상적 가치, 관련성과 구체성**, 그리고 일반 상식에만 기대지 않고 질문에 직접 답하는지

시각 선호쌍 Dv → visual tool confidence sv

- MedKLIP의 **병변 검출 confidence를 그대로 그 샘플의 가중치로 사용**한다
- 논리: 노이즈를 씌운 영역이 부정확하면 그 샘플의 신뢰도가 낮으므로 **학습에 덜 반영**해야 한다
- 즉 여기서는 "답이 얼마나 좋은가"가 아니라 **"병변을 제대로 찾았는가"**가 가중치가 된다

정직 주석 - heatmap을 스칼라 점수로 바꾸는 구체적 산식은 **논문에 명시되지 않았다** (알고리즘에 $s \leftarrow P(h|x)$ 라고만 표기).

두 점수는 **하나로 합쳐지지 않는다.** 각 샘플이 자기 종류의 점수를 가진 채로 합집합 D에 들어가고, 거기서 함께 정규화된다.

최종 목적함수 - 가중 DPO

배경 - 표준 DPO (논문 식 1)

$$L_{DPO} = -E[\log \sigma(\alpha \cdot \log(\pi_\theta(y_w|x) / \pi_{ref}(y_w|x)) - \alpha \cdot \log(\pi_\theta(y_l|x) / \pi_{ref}(y_l|x)))]$$

읽는 법 - 좋은 답의 **상대 확률**을 올리고 나쁜 답의 상대 확률을 내린다. π는 SFT를 마친 고정 참조 모델, α는 온도.

MMedPO (논문 식 3)

$$L_{MMedPO} = -E[s' \cdot \log \sigma(\alpha \cdot \log(\pi_\theta(y_w|x) / \pi_o(y_w|x)) - \alpha \cdot \log(\pi_\theta(y_l|x^*) / \pi_o(y_l|x^*)))]$$

차이 ① 정규화 점수 s′ 가 log σ(·) **전체에 곱해지는 샘플별 가중치**다 (logit 안이 아니라 바깥). 정규화는 $s' = (s - \mu)/\sigma$ 후 [α, β]로 clip, 부록 값 α=0.75 · β=1.25.

차이 ② ★ dispreferred 항의 조건이 x가 아니라 **x* (병변이 가려진 이미지)** 다.

시각 선호쌍 Dv에서는 $y_w = y_l$ = 정답 y 이므로, 위 식이 이렇게 된다:

$$L = -s' \cdot \log \sigma(\alpha \cdot \log(\pi_\theta(y|x) / \pi_o(y|x)) - \alpha \cdot \log(\pi_\theta(y|x^*) / \pi_o(y|x^*)))$$

→ **"병변이 보일 때의 정답 확률이, 병변이 가려졌을 때보다 높아야 한다"**를 직접 최적화한다.

모델이 텍스트 사전지식만으로 y를 만든다면 두 항이 같아져 log σ(0)에서 **손실이 전혀 줄지 않는다.**
텍스트 지배 자체가 손실의 원천이 되는 구조

실험 세팅

학습 자체는 가볍다 - **LoRA로 A6000 4장에 2–3시간.** 비용은 데이터 큐레이션(GPT-4o · Med-LLM 3종) 쪽에 있는데, 그 비용은 논문에 보고되지 않았다.

학습

항목	값
베이스 모델	LLaVA-Med-1.5 7B
호환성 검증용 2번째 백본	LLaVA-Med++
미세조정	LoRA
learning rate / batch / epochs	1e-7 / 4 / 3
GPU · 시간	4 × RTX A6000 · 약 2–3시간
Visual tool	MedKLIP (프롬프트 "disease")
Med-LLM 3종	Med42-7B, Med42-70B, BioMistral-7B (5라운드)
dispreference 생성기	GPT-4o

학습 데이터 - 시각쌍 : 텍스트쌍 = 정확히 1:1

데이터셋	시각쌍	텍스트쌍	합
SLAKE	4,919	4,919	9,838
IU-Xray	2,069	2,069	4,138
VQA-RAD	1,797	1,797	3,594
MIMIC-CXR	800	800	1,600

평가 지표

Med-VQA (SLAKE, VQA-RAD) - open-ended는 **recall**, closed-ended는 **accuracy**

Report generation (IU-Xray, MIMIC-CXR) - **BLEU(1–4 평균), ROUGE-L, METEOR**

비교 baseline 6종

DPO · Self-Rewarding · POVID · SIMA · FiSAO · STLLaVA-Med - 각각에 대해 SFT를 먼저 적용한 조건도 함께 보고한다. POVID는 "환각 주입 + 이미지 왜곡"을 쓰는, 구조가 가장 비슷한 선행 연구다.

결과 - 4개 데이터셋에서 전 baseline 상회

LLaVA-Med v1.5 기반, SFT 미적용 조건. VQA는 open=recall / closed=accuracy, 리포트는 BLEU(1–4 평균)·ROUGE-L. **굵은 빨강 = 최고값.**

방법	SLAKE Open	SLAKE Closed	VQA-RAD Open	VQA-RAD Closed	IU-Xray BLEU	IU-Xray ROUGE-L	MIMIC BLEU	MIMIC ROUGE-L
LLaVA-Med v1.5 (원본)	44.26	61.30	29.24	63.97	14.56	10.31	10.25	9.38
+ DPO	49.30	62.02	29.76	64.70	16.08	12.95	11.19	9.45
+ POVID	52.43	70.35	31.77	65.07	20.80	24.33	11.21	9.66
+ FiSAO	52.69	70.46	32.70	64.11	21.06	25.72	11.32	9.68
+ SIMA	51.77	69.10	31.23	64.80	17.11	22.87	11.16	9.58
+ MMedPO (제안)	**53.99**	**73.08**	**36.36**	**66.54**	**23.49**	**29.52**	**12.85**	**11.13**

출처: Zhu & Xia et al., MMedPO, ICML 2025 (arXiv:2412.06141), Table 1

Ablation - 어느 부품이 실제로 일했나

세 가지를 떼어 본다. 결론부터 말하면 **향상의 거의 전부가 시각 대조(Stage 2)에서 나온다.**

① 두 전략 분해 (IU-Xray BLEU-1)

Stage 1(환각 주입)만 43.95
Stage 2(병변 노이즈)만 **55.15**
Stage 1+2 **55.58**

Stage 2만으로 이미 11.2pt를 벌고, 결합으로 더해지는 것은 **0.43pt**뿐이다. 단 Med-VQA에서는 두 전략의 기여가 서로 비슷하다고 논문이 밝힌다.

② 임상 관련성 점수(CRS) 유무

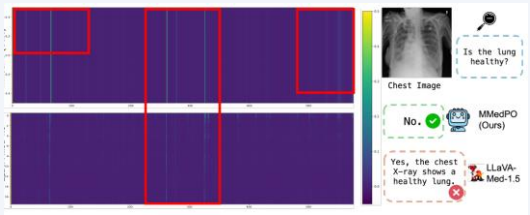
Med-VQA 평균 **+2.3%**
리포트 생성 평균 **+18.5%**
IU-Xray 19.30 → **25.00**

18.5%는 IU-Xray 한 데이터셋이 끌어올린 값이다. MIMIC-CXR은 Stage 1에서 6.55 → 6.58, Stage 2에서 7.17 → 7.24로 **사실상 무변화**다.

③ 국소 vs 전역 노이즈

SLAKE 58.88 → **59.88**
VQA-RAD 46.91 → 46.98
IU-Xray 24.88 → 25.00
MIMIC 6.80 → 7.24

방향은 4개 데이터셋에서 일관되나 마진이 **0.07–1.0pt**다. 시드 반복도 신뢰구간도 없어 "일관되게 우수하다"는 서술은 과하다.



위=MMedPO, 아래=LLaVA-Med-1.5의 image token attention. 빨간 박스가 강화된 영역.

정성 사례 - 텍스트 사전지식이 이기는 순간

복부 MRI, "이 영상에 신장이 몇 개인가?" → 정답 **0** · LLaVA-Med "두 개" · DPO "2" · **MMedPO "0"**

영상에 신장이 나오지 않는데도 baseline은 "복부 영상이면 신장이 두 개"라는 언어 사전지식으로 답한다. 처방이 겨냥한 실패 모드를 정확히 보여 주는 사례

출처: Zhu & Xia et al., MMedPO, ICML 2025 (arXiv:2412.06141), Table 2 · Table 4 · Table 8 · Fig. 5 · Fig. 10

Limitations - 메커니즘은 명확한데 증거가 비어 있다

이 논문에는 별도의 Limitations 절이 없음. 제 개인적인 생각입니다.

측정의 공백 - 목표와 증거가 어긋난다

- **"시각 의존도가 실제로 올랐다"는 직접 측정이 없다.** 이미지를 가리거나 다른 이미지로 바꿔 넣는 대조 실험이 논문에 없다
- 있는 증거는 둘뿐 - **attention map 1장(정성)**, 그리고 국소 vs 전역 노이즈 비교인데 후자는 간접이고 **마진이 0.07–1.0pt**로 매우 작다
- **환각 전용 벤치마크(POPE / CHAIR / CARES 등) 평가가 없다.** 저자 그룹이 CARES를 직접 만들었는데도 쓰지 않았다
- 리포트 평가가 BLEU · ROUGE-L · METEOR 같은 **n-gram 지표뿐**이다. RadGraph F1 같은 임상 사실성 지표를 쓰지 않았다

보고 · 재현성 · 통계

- 헤드라인 수치(초록 **14.2% / 51.7%**)의 **기준선 서술이 초록·서론·§4.2에서 서로 다르고** Table 1에서 재현되지 않는다
- 노이즈 단계 **k 값과 ξ 스케줄이 미명시**이고, heatmap을 스칼라 점수로 바꾸는 산식도 없다 → 논문만으로는 재현 불가
- **시드·반복 횟수 언급이 없고 신뢰구간·유의성 검정도 없다.** 평가셋도 작다 (MIMIC-CXR 리포트 200건, VQA-RAD 451문항)
- 임상 관련성 점수의 이득이 **IU-Xray에 편중**된다 - MIMIC-CXR은 6.55 → 6.58로 사실상 무변화

내 생각: 처방의 **메커니즘은 손실함수 안에 분명히 있는데**, 그 처방이 통했는지를 재는 지표가 최종 성능밖에 없다. 즉 **이 논문은 자기 주장의 핵심을 직접 측정하지 않는다.**

무엇이 얼마나 부족한가 - 능력을 쪼개서 재기

정확도 한 숫자로는 병목을 알 수 없다. Med-CMR은 **시각 이해와 임상 추론을 7개 축으로 분해**해 18개 모델을 stress test 한다.

Med-CMR: A Fine-Grained Benchmark Integrating Visual Evidence and Clinical Logic for Medical Complex Multimodal Reasoning

Haozhen Gong[1,†] Xiaozhong Ji[2,†] Yuansen Liu[1,†] Wenbin Wu[1] Xiaoxiao Yan[4] Jingjing Liu[1]
Kai Wu[3] Jiazhen Pan[5] Bailiang Jian[5] Jiangning Zhang[6] Xiaobin Hu[1,*] Hongwei Bran Li[1]
[1]National University of Singapore [2]Nanjing University [3]Tongji University [4]Ruijin Hospital
[5]Technical University of Munich [6]Zhejiang University
† Equal contribution * Corresponding author

arXiv 2512.00818 (v1 2025-11 / v2 2026-03)
Haozhen Gong, Xiaozhong Ji, Yuansen Liu 외 12인 (공동 1저자 3명)
싱가포르국립대 · 난징대 · 통지대 · 루이진병원 · 뮌헨 공대 · 저장대
투고 venue 미명시 · peer review 이전
데이터는 HuggingFace 공개, 평가 코드는 미공개

이 논문이 한 일

- 의학 저널 증례보고에서 **VQA 20,653문항**을 새로 구축 (기존 VQA 재활용 아님)
- 능력을 **시각 복잡도 3축 + 추론 복잡도 4축**으로 분해해 축마다 따로 채점
- **18개 MLLM** 평가 - 상용 2 · 오픈 범용 12 · 의료 특화 4
- base 모델과 그 **의료 특화 버전을 쌍으로** 직접 비교

왜 새 벤치마크가 필요한가

기존 의료 VQA 벤치마크는 크거나(커버리지) 어렵거나 둘 중 하나였고, **능력을 축별로 나눠 채점한 것은 없었다**.

벤치마크	문항 수	QA 유형	넓은 커버리지	어려운 태스크	세분화 평가
VQA-RAD	451	Open	×	×	×
Path-VQA	6,719	Open	×	×	×
Slake-En	1,061	Open	×	×	×
PMC-VQA	33,430	Open	✓	×	×
OmniMedVQA	127,995	MCQ	✓	×	×
GMAI-MMBench	21,281	MCQ+MRQ	✓	×	×
MedXpertQA MM	2,000	MCQ	✓	✓	×
HIE-Reasoning	749	MCQ+Open	×	✓	×
Med-CMR (본 논문)	**20,653**	**MCQ+Open**	✓	✓	✓

한계 1 - 인지 수준에 머문다

대부분 **perception 수준 VQA**다. 이미지를 묘사하거나 명백한 사실을 검색하는 수준이라, 실제 임상 판단을 가르는 어려운 케이스가 빠져 있다.

한계 2 - 단일 점수로 뭉뚱그린다

"의료 멀티모달 추론"을 **점수 하나로 취급**한다. 시각 이해(보기)와 임상 추론(생각하기)을 분리해서 재지 않는다.

결과 - 병목을 특정할 수 없다

점수가 낮을 때 **못 본 것인지 못 따진 것인지 알 수 없다.** 고치려면 먼저 나눠서 재야 한다는 것이 이 논문의 출발점이다.

출처: Gong et al., arXiv:2512.00818, Table 1 발췌

Contributions - 이 논문이 내놓은 네 가지

Med-CMR이 내놓은 것은 새 모델이 아니라 **새로운 자**다. 무엇을 어떻게 재는지를 네 갈래로 정리한다.

i

능력을 7개 축으로 분해해 축마다 따로 채점한다

시각 복잡도 3축(SOD · FDD · SU)과 추론 복잡도 4축(TP · CR · LTG · MSI)으로 나눈 **최초의 세분화 의료 멀티모달 벤치마크**다. 점수가 하나면 못 본 것과 못 따진 것이 구분되지 않는다.

근거 slide 45–46

ii

기존 VQA 재활용이 아니라 문헌 원문에서 새로 만들었다

의학 저널 증례보고에서 **이미지와 사람이 쓴 캡션을 직접 가져와 20,653문항**을 구축했다. 5지선다 16,655개, 자유서술형 3,998개다.

근거 slide 47–48

iii

오답을 시각 의존적으로 설계하고 사람이 여섯 번 개입한다

오답 4개가 **텍스트만이 아니라 시각 정보 해석에 의존해야** 풀리도록 기준을 걸었다. 템플릿 설계부터 면허 의사 전수 리뷰까지 사람이 여섯 번 거른다.

근거 slide 47

iv

18개 MLLM을 stress test 해 병목을 시각 쪽으로 지목했다

언어 축은 만점 근처(GPT-5 명확성 **97.77**)인데 **시각 정확도는 40.45**다. 게다가 13개 지표 중 시각 정확도만 모델 크기와의 상관이 가장 약하다(**r=0.59**).

근거 slide 49

능력 분해 ① - 시각 복잡도 3축

"이미지를 얼마나 잘 보는가"를 세 가지로 나눈다. 아래 예시는 모두 **논문 Fig. 1에 실제로 실린 문항**이다.

SOD

작은 이상 소견 탐지

Small-Object Detection

몇 픽셀에 불과한 **작고 저대비인** 이상 소견을 찾아낸다.

예시 · 내시경 영상

"겉모습으로 볼 때, 이 아주 작은 구조물이 진짜 병변임을 뒷받침하는 소견은?"

정답: Dieulafoy 유사 병변에 의한 대량 출혈

점막 아래 비정상 혈관이 터져 대량 출혈을 일으키는 드문 병변. 몇 mm 크기라 놓치기 쉽다.

FDD

미세 차이 판별

Fine-Detail Discrimination

겉보기는 비슷하지만 **임상 의미가 다른** 소견을 구별한다.

예시 문항

"이 분포 양상이 가장 잘 나타내는 것은?"

정답: 다발성처럼 보이지만 실제로는 한 곳에서 점막 아래로 퍼진 것

병변이 하나인지 여럿인지에 따라 수술 절제 범위가 달라진다.

SU

공간 이해

Spatial Understanding

여러 장·여러 모달리티의 영상을 결합해 **공간적 일관성**을 유지한다.

예시 · 척추 3D 영상

"이 분포는 인접 구조물과의 공간적 관계에 대해 무엇을 말해 주는가?"

정답: 흉추 5–7번에 걸쳐 연속된 병변

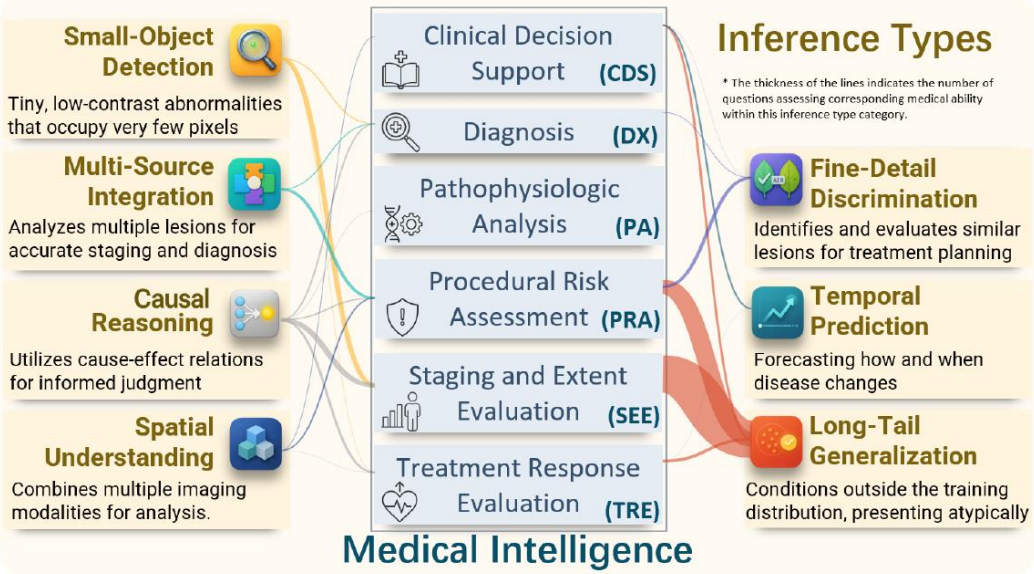
떨어져 있는지 이어져 있는지를 슬라이스를 넘나들며 추적해야 답이 나온다.

세 축 모두 "보는 것" 자체가 어려운 문제다 - 추론을 아무리 잘해도, 못 본 것은 만회되지 않는다.

출처: Gong et al., arXiv:2512.00818, Fig. 1 예시 문항 (한국어 요약)

능력 분해 ② - 추론 복잡도 4축

"보고 나서 어떻게 따지는가"를 네 시나리오로 나눈다. 7개 축 각각에 **따로 점수를 낸다** - 이것이 fine-grained의 뜻이다.



시간 추론
TP · Temporal Prediction
질병이 **언제 어떻게 변할지**, 진행과 예후를 추론한다.

인과 추론
CR · Causal Reasoning
증상 · 소견 · 결과를 잇는 **다단계 인과 사슬**을 따라간다.

희귀 사례 일반화
LTG · Long-Tail Generalization
거의 못 본 희귀 질환, 교과서와 다르게 나타나는 케이스.

다중 근거 통합
MSI · Multi-Source Integration
여러 이상 소견에서 핵심 단서를 뽑아 **병기**(병이 퍼진 정도)를 매긴다.

같은 문항을 "의료 지능" 6축으로도 다시 분류한다

CDS 임상 의사결정 지원 · DX 진단 · PA 병태생리 분석 · PRA 시술 위험 평가 · SEE 병기 · 범위 평가 · TRE 치료 반응 평가 (그림 가운데 열 - 선 두께가 해당 문항 수)

출처: Gong et al., arXiv:2512.00818, Fig. 2 좌측 패널

데이터 구축 - 문헌 원문에서 새로 만든다

기존 VQA 데이터셋을 재활용하지 않고, **의학 저널의 증례보고에서 이미지와 사람이 쓴 캡션을 직접 가져와** 문항을 새로 만들었다.

1. 수집

의학 저널의 **증례보고 · 연구논문**(예: JMCR, NEJM)에서 이미지 + 사람이 쓴 캡션 + 메타데이터를 수집한다.

기존 VQA 재활용 아님

→

2. 문항 생성

의학 배경 annotator가 카테고리별 **10–20개 템플릿**을 설계. **GPT-5-mini**가 이미지마다 템플릿을 고르고 캡션에서 정답을 추출한다.

→

3. 오답 선택

MLLM 3종(GPT-5-Mini · Qwen3-VL-Plus · Claude-Sonnet-4)이 각 4개씩 **총 12개 후보** 생성 → 의학 배경 annotator 3명이 **4개 선정**.

→

4. 필터링 2단계

① 사람이 이미지를 전수 리뷰(캡션 부실 · 차원 불일치 제거) ② Lingshu-7B · Qwen2.5-VL-7B · LLaVA-Med-1.5 **3개가 모두 맞힌 문항은 전부 제거**.

→

5. 품질 보증

합의 리뷰(2명 + 독립 auditor 1명) → 정답 유일성 검증(4명) → **면허 의사 전수 리뷰**. 원 합성 데이터의 **8%를 제거**했다.

오답 4개가 만족해야 할 기준 (3단계)

- ① 충분한 난이도 - 잘 준비된 응시자에게도 어려울 것
- ② 확실히 틀릴 것 - 정답과 의미가 겹치지 않을 것
- ③ **텍스트만이 아니라 시각 정보 해석에 의존할 것**
- ④ 임상적으로 그럴듯할 것

정직하게 짚을 것

③이 텍스트 지배를 겨냥한 **유일한 설계 장치**다. 다만 그것이 실제로 작동했는지는 **사후에 측정되지 않았다** - 이미지를 가린 조건에서 정답률이 얼마나 떨어지는지를 재지 않았다.

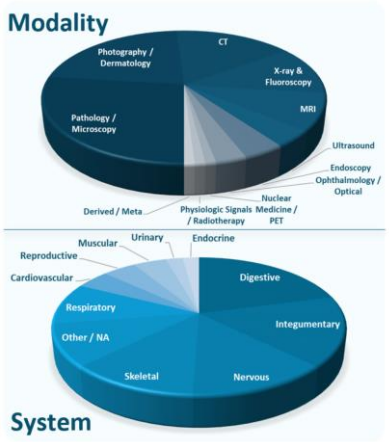
이미지를 빼고 재 본 것은 논문 ①뿐이다.

사람이 여섯 번 개입한다 - 템플릿 설계 → 오답 4개 선정(3명) → 이미지 전수 스크리닝 → 합의 리뷰(2명 + auditor 1명) → 정답 유일성 검증(4명) → 면허 의사 전수 리뷰. LLM이 초안을 만들고 사람이 걸러 내는 구조다.

출처: Gong et al., arXiv:2512.00818, §2.2

무엇으로 이루어져 있나

총 **20,653개 VQA** - MCQ 16,655 (80.6%) · open-ended 3,998 (19.4%). 장기계 11종 × 영상 모달리티 12종.



영상 모달리티 (상위 7종)	비율
병리 / 현미경	**26.4%**
사진 / 피부과	**21.8%**
CT	16.1%
X-ray & 투시	12.1%
MRI	9.4%
초음파	4.2%
내시경	3.6%

장기계 (상위 6종)	비율
소화기	18.6%
피부	15.6%
신경	12.3%
골격	11.0%
호흡기	8.8%
심혈관	8.8%

관전 포인트

방사선 영상(CT + X-ray + MRI + 초음파 + 핵의학) 43.8%보다 **병리 + 피부 사진이 48.2%로 더 크다.**

증례보고에서 뽑았기 때문에 생기는 선택 편향이다 - 보고할 만큼 희귀하고 사진이 인상적인 케이스. 논문은 이를 논하지 않는다.

평가 프로토콜

- MCQ = 정규표현식으로 보기 문자(A–E)를 뽑아 정확도 계산
- Open-ended = **DeepSeek-V3.2-Exp를 judge로** 4차원 가중 채점 - 명확성 10% · 추론 일관성 10% · **시각 정확도 40%** · **정답 정합 40%**, 0–100으로 정규화
- ⚠ **judge는 이미지를 보지 않는다** - "시각 정확도"를 사람이 쓴 이미지 캡션 **텍스트**와 대조해 채점한다 (이 논문 최대 주장의 측정 타당성을 흔드는 지점...)

출처: Gong et al., arXiv:2512.00818, Fig. 2 우측 패널 · Table 6 · Table 7 · §2.4 | 파이는 구성 비중만, 세부 수치는 오른쪽 표

결과 - 최고 모델도 MCQ 57.8%

논문 인용: "유창하고 내적으로 일관된 추론은 **비교적 성숙**한 반면, **올바른 시각 증거를 추출해 정답으로 수렴하는 것이 주된 병목으로 남아 있다.**"

모델	SOD	FDD	SU	TP	CR	LTG	MSI	MCQ All	명확성	일관성	시각 정확도	정답 정합	Open All
GPT-5	66.08	71.45	62.06	58.33	60.30	**55.19**	69.00	**57.81**	**97.77**	88.86	**40.45**	**34.65**	**48.70**
Gemini-2.5-Pro	58.75	68.07	56.70	52.08	53.54	46.42	64.42	**49.87**	98.11	89.36	35.77	32.07	**45.98**
Qwen3-VL-235B-A22B	57.48	66.95	55.99	55.06	53.33	45.86	63.07	**49.34**	97.10	85.21	33.02	27.95	**42.62**
InternVL3.5-241B-A28B	55.91	65.68	52.47	54.17	48.80	42.73	56.33	**46.17**	96.87	76.33	**42.74**	33.66	**47.88**
Qwen2.5-VL-72B	52.10	61.32	47.39	51.19	46.36	38.46	54.18	**42.17**	96.31	75.11	33.29	25.68	**40.73**
Gemma3-27B	45.94	52.74	45.98	37.80	42.35	33.62	45.28	**37.07**	93.72	76.23	25.96	19.96	**35.36**
Medgemma-27B 의료 특화	37.44	47.54	37.24	29.46	30.80	25.92	36.93	**28.91**	89.82	72.21	20.64	17.58	**31.49**
Medgemma-4B 의료 특화	16.13	17.72	13.12	14.58	17.64	14.00	23.45	**14.90**	86.98	57.17	26.65	17.57	**32.10**

언어는 만점 근처, 시각은 바닥 - 같은 모델 안에서

97.77 GPT-5 · 명확성 **VS** **40.45** GPT-5 · 시각 정확도

전 모델을 통틀어 **시각 정확도 최고값이 42.74** - 절반을 넘는 모델이 하나도 없다.

Long-Tail Generalization이 7축 중 최저

희귀 · 비전형 케이스 축은 **모든 모델에서 최저**였다. 최고가 GPT-5 **55.19%**이고 **오픈소스는 전부 46% 미만**이다. GPT-5 자신의 최고 축(FDD 71.45)과도 16점 넘게 벌어진다.

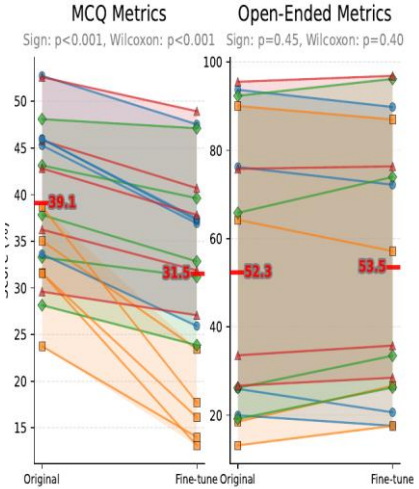
우연보다 못한 모델도 있다

5지선다이므로 **우연 정답률은 20%**. Medgemma-4B의 **14.90%**는 그보다 낮다. 모델 크기와의 상관도 시각 정확도가 **r=0.59**로 13개 지표 중 가장 약하다.

출처: Gong et al., arXiv:2512.00818, Table 2 (18개 모델 중 8개 발췌) · Fig. 3

반전 - 의료 파인튜닝이 시각 근거를 약화시킨다

같은 base 모델과 그 의료 특화 버전 **4쌍**을 직접 짝지어 비교했다. 도메인 데이터를 더 넣는 것이 왜 손해가 될 수 있는지가 이 절의 주제다.



base → 의료 특화 (MCQ All)	base	의료 특화	변화
Gemma3-4B → Medgemma-4B	25.98	14.90	**-11.08**
Gemma3-27B → Medgemma-27B	37.07	28.91	**-8.16**
Qwen2.5-VL-7B → Lingshu-7B	31.06	27.26	**-3.80**
Qwen2.5-VL-32B → Lingshu-32B	33.36	30.47	**-2.89**

왜 그런가 - 틀린 문항 300개를 사람이 직접 읽었다

"올바른 **진단 방향은 잡았으나 미세한 시각 단서를 놓쳐** 오답에 도달한다" · "몇 개의 두드러진 특징을 **전형적 진단에 곧바로 연결하는 pattern matching**에 자주 의존하며, 다른 맥락 정보를 통합하지 않는다"

GPT-5의 오류를 7축 × 각 100건 수동 분류했을 때도 **7축 중 6축에서 "이미지 인식" 오류가 최대 비중**이었다. 두 관찰이 같은 곳을 가리킨다.

→ 의료 파인튜닝이 도메인 연상 패턴 의존을 늘리고 교차모달 증거 통합을 줄인다

4쌍 평균과 유의성

MCQ **39.1 → 31.5** · Sign · Wilcoxon 모두 **p<0.001**

Open-ended 52.3 → 53.5 · p=0.45 / 0.40 **유의하지 않음**

다만 방향이 일관되지는 않는다

자유서술형 평균은 올랐지만 유의하지 않았고, 방향도 모델마다 갈린다. 논문은 일반 모델에 유리하던 MCQ **500개**를 자유서술형으로 바꿔 다시 채점했다. Lingshu-32B는 base인 Qwen2.5-VL-32B를 **역전**했지만, Medgemma-27B는 base인 Gemma3-27B에 **여전히 뒤졌다**.

→ 의료 파인튜닝이 얻는 것(도메인 어휘와 서술)과 잃는 것(미세 시각 판별)의 균형이 모델마다 다르다

출처: Gong et al., arXiv:2512.00818, Fig. 4(b)(c) · Table 2 · §4.2

Limitations - 이 자로는 아직 못 재는 것

벤치마크 논문의 한계는 **측정 도구 자체의 한계**다. 특히 앞의 셋은 이 논문의 최대 주장을 직접 겨눈다.

측정 설계의 공백

- **인간 · 전문가 기준선이 없다** - 절 제목이 "How Far Are We from Expert-Level"인데 정작 expert 수치가 하나도 없다. 57.81%가 잘한 것인지 못한 것인지 판단할 기준점이 없다
- **blind(이미지 없는) 대조군이 없다** - 시각 의존성을 오답 선정 기준으로만 걸고 사후에 측정하지 않았다. 5지선다이므로 우연 정답률 20%와 대조했어야 한다
- **judge가 이미지를 보지 않는다** - DeepSeek-V3.2-Exp가 사람이 쓴 캡션 텍스트와 대조해 "시각 정확도"를 매긴다. 이 논문 최대 주장의 측정 타당성을 흔드는 지점이다

데이터 · 재현성

- **자기 오염 위험** - 오답 생성에 GPT-5-mini · Qwen3-VL-Plus · Claude-Sonnet-4가 쓰였는데 평가 1위가 GPT-5, 오픈소스 1위가 Qwen3-VL 계열이다. 최종 선택은 사람이 했고, 오염 점검은 ROUGE-L·편집거리 같은 **어휘 중복**뿐이다
- **선택 편향** - 증례보고 출처라 보고할 만큼 희귀하고 사진이 인상적인 케이스에 치우친다. 병리 26.4% + 피부 21.8% = 48.2%가 방사선 전체 43.8%보다 크다. 논문은 이를 논하지 않는다
- **재현 정보 부족** - 평가 코드 미공개, 모델 스냅샷과 reasoning effort 미기재
- **arXiv 프리프린트** - 투고 venue 미명시, 아직 peer review를 거치지 않았다

정리하면 **"시각이 병목"이라는 결론 자체는 그럴듯하지만, 그것을 재는 도구가 그 결론을 검증하도록 만들어져 있지 않다.** 축별 분해라는 아이디어와, 그 축을 실제로 재는 방법은 분리해서 평가해야 한다.

개인 의견 - 문제의식은 맞고, 측정이 못 따라갔다

높이 사는 점

"점수 하나로 뭉뚱그리지 않는다"는 문제의식이 정확하다

근거: 점수가 낮을 때 **못 본 것인지 못 따진 것인지** 구별되지 않으면 무엇을 고쳐야 할지 정할 수 없다. 축별 분해는 그 구별을 가능하게 하는 최소 장치다.

base와 의료 특화를 직접 짝지어 비교했다

근거: 4쌍 모두 5지선다에서 하락했고 **Sign · Wilcoxon 모두 $p<0.001$**이다. "의료 데이터로 더 학습하면 낫다"는 이 분야의 기본 가정에 대한 실질적 경고다.

납득하기 어려운 점

가장 중요한 주장을 이미지를 안 보는 judge로 채점했다

근거: 시각이 병목이라는 결론이 **캡션 텍스트 대조로 매긴 "시각 정확도"**에 걸려 있다. 측정 도구가 주장을 지지하지 못한다.

blind 대조군을 한 번도 돌리지 않았다

근거: 이미지를 가리고 같은 문항을 풀리면 설계 전제를 바로 검증할 수 있다. **추가 학습도 새 데이터도 필요 없는**, 비용이 거의 들지 않는 실험이다.

그래서 어떻게 쓸 것인가

그럼에도 **능력 분해 축 자체는 재사용 가치가 크다.** 데이터가 HuggingFace에 공개돼 있으므로, 빠진 blind ablation은 우리가 직접 돌려 볼 수 있다.

이미지를 빼고 다시 풀린 뒤 **우연 정답률 20%**와 비교하면 된다.

세 논문 잇기 - 맞물리는 곳과 어긋나는 곳

세 논문은 서로를 인용하지 않으며, 같은 데이터·같은 모델을 쓰지도 않았다.

맞물리는 곳 - 서로 다른 방법으로 같은 결론에 도달한다

- ① **행동 수준** 모달리티를 지우고 텍스트를 교란해 **텍스트 지배를 인과적으로 입증**했다 - 이미지만으로 맞히던 문제가 텍스트 하나로 무너진다
- ② **개입 수준** 시각 대조 손실을 넣었더니 **성능 향상의 대부분이 시각 쪽(Stage 2)에서 나왔다** - 텍스트 환각 대조(Stage 1)보다 훨씬 크다
- ③ **계량 수준** 언어 축은 만점 근처(97.8)인데 **시각 축은 40.5**, 게다가 **시각 정확도만 모델 크기와의 상관이 최약(r=0.59)** - 스케일링으로도 안 메워진다

어긋나는 곳 - 이 세 편을 그대로 이어 붙이면 안 되는 이유

- **처방과 측정이 정면으로 충돌한다.** ②는 의료 도메인 정렬로 고치려 하는데, ③은 의료 특화 파인튜닝이 MCQ에서 **4쌍 모두 하락(p<0.001)**했다고 보고한다. 다만 ③이 본 것은 대규모 도메인 SFT(Medgemma·Lingshu)이고 ②는 시각 대조 선호 최적화라 종류가 다르다 - 반증은 아니지만 경고다.
- **"인간은 텍스트로 좋아지지 않는다"는 검정력이 약하다.** 의사 3명, 전원 공저자, 96문항. 긴 텍스트 구간의 신뢰구간은 0.297–0.789로 매우 넓다. 이것은 **증거의 부재**에 가깝지 **부재의 증거**가 아니다.
- **①의 오도 실험은 설계상 최대 적대적이다.** vignette를 "영상·검사·신체진찰 묘사 금지"로 생성해 텍스트와 이미지가 완전히 독립적으로 다른 병을 가리키게 만들었다. 실제 임상에서 이런 불일치가 얼마나 흔한지는 논문이 제시하지 않는다.
- **②③ 모두 자기 주장을 자기가 측정하지 않았다.** ②는 "시각 의존도가 올랐다"를 재지 않았고, ③은 "시각이 병목이다"를 **이미지를 안 보는 judge**로 채점했다.

남은 공백 - 세 편 모두가 비켜 간 것

세 논문의 한계를 겹쳐 놓으면 **아무도 하지 않은 실험 네 개**가 남는다.

GAP 1 · 가장 값싸고 가장 중요

처방의 효과를 시각 의존도로 재기

"정렬했더니 BLEU가 올랐다"는 **시각 근거를 더 썼다는 증거가 아니다.** 필요한 것은 정렬 전후에 **이미지를 가리거나 다른 이미지로 바꿔 넣고** 성능 낙폭을 비교하는 것 - 낙폭이 커졌으면 이미지에 더 의존하게 된 것이다.

② MMedPO: 미측정 (attention map 1장뿐) · ③ Med-CMR: blind 대조군 없음

GAP 2 · 기전

인코더가 약한가, LLM이 무시하는가

둘은 완전히 다른 처방을 요구한다. 행동 수준 증거만으로는 구별되지 않는다. 실제로 같은 945건에서 **o3의 이미지 전용 정확도는 64.9%, GPT-4V Turbo는 37%**였다 - 인코더 능력 차이만으로도 상당 부분 설명된다.

① 논문이 자인한 한계: "내부 기전을 연구하지 않았다"

GAP 3 · 추론 시점 완화

"이미지 먼저" 강제 프로토콜

텍스트를 보기 전에 **이미지만으로 먼저 소견을 쓰게 하고**, 그다음 텍스트를 주는 2단계 프롬프트. 학습 없이 즉시 시험 가능하고, 오도 텍스트 실험에 그대로 적용된다.

① 논문이 명시적으로 future work로 남긴 것 - 아무도 아직 하지 않았다

GAP 4 · 비교의 대칭성

인간에게도 같은 절제를 시키기

모델만 이미지 전용 / 텍스트 전용으로 쪼개고, **의사는 멀티모달 한 조건만** 풀었다. "인간은 텍스트에 덜 기댄다"를 주장하려면 인간에게도 같은 세 조건을 줘야 한다.

① 설계상 비대칭 - 결론의 방향은 그럴듯하나 근거가 대등하지 않다

우리가 가져갈 것

방법론 원칙 하나와, **지금 당장 돌릴 수 있는 실험 세 개**.

원칙 - 멀티모달 결과를 보고할 때는 blind 대조군을 함께 낸다

- 모달리티를 하나 지운 조건의 점수를 **같은 표에** 보고하면, "이 성능이 어느 모달리티에서 왔는가"가 한눈에 보인다
- 비용이 거의 들지 않는다 - 같은 프롬프트에서 이미지만 빼고 한 번 더 돌리면 된다
- 이것을 안 하면, 우리 결과도 오늘 본 세 논문과 똑같은 지적을 받는다

즉시 가능한 실험

- **Med-CMR blind ablation.** MCQ 16,655문항이 HuggingFace에 공개돼 있다. 이미지를 빼고 같은 모델을 돌린다. 5지선다이므로 **우연 정답률 20%**. GPT-5의 57.81%가 blind에서 몇 %인지가, 이 벤치마크의 "visual dependence" 설계가 실제로 작동했는지를 판정한다.
- **"이미지 먼저" 2단계 프롬프트.** 이미지만으로 소견을 먼저 쓰게 한 뒤 텍스트를 준다. 논문 ①의 오도 vignette는 **348건 × 오답 4개 = 1392 trial의 원문이 GitHub에 공개**돼 있다. 단, 원본 이미지는 레포에 없어 NEJM에서 따로 구해야 한다 - 텍스트 조건만 즉시 재현 가능.
- **정렬 전후 시각 의존도 델타.** MMedPO 체크포인트에 이미지 가림·치환을 적용해 낙폭을 재면, 논문 ②가 재지 않은 것을 바로 메울 수 있다.

우리 프로젝트와의 접점 (추측)

- VLM speculative decoding에서 draft 모델은 **작고 언어 편향이 더 강하다**. 텍스트 prior만으로 그럴듯한 토큰을 뽑으면 **수용률은 높은데 시각 근거는 없는** 상태가 될 수 있다.
- 검증 방법은 오늘 본 것과 같다 - 이미지를 가렸을 때 **수용 길이가 얼마나 떨어지는가**를 재면, drafter가 시각 신호를 실제로 쓰는지 알 수 있다. 떨어지지 않으면 drafter는 이미지를 안 보고 있는 것이다.
- 단, 이것은 오늘 세 논문이 다룬 내용이 아니라 **우리가 확인해 봐야 할 가설**이다.

정리하며

이미지를 넣었다고 해서, **이미지를 본 것은 아니다**

① 진단

o3는 텍스트만으로 **66.0%**를 맞혔고, 텍스트가 길어질수록 이미지의 기여는 사라졌다. 오도하는 서술 하나에 **84% → 28%**.

② 처방

이미지를 봐야만 구별되는 선호쌍을 만들면 개선된다. 다만 **"시각 의존도가 올랐다"는 측정되지 않았다.**

③ 측정

언어 축 **97.8** 대 시각 축 **40.5**. 스케일링과의 상관도 시각 축만 최약 (r=0.59)이다.

멀티모달 성능을 보고할 때 **모달리티를 지운 조건**을 함께 내자.

한 줄 더 돌리는 비용으로, 우리 결과가 어디에서 왔는지 스스로 답할 수 있게 된다.

감사합니다.