3. LLM을 어떻게, 왜 사용했는지 (방법론)

연구진은 **MLLM-SRec**이라는 새로운 프레임워크를 제안했음.

(1) 왜 LLM을 사용했는가

- LLM은 **세계 지식**과 **언어적 추론 능력**을 갖추고 있어, 텍스트 기반 추천 성능을 강화할 수 있음.
- 멀티모달 LLM(MLLM)은 **이미지+텍스트 융합**을 통해 기존 멀티모달 추천의 한계를 보완할 수 있음.
- CoT reasoning으로 **동적 사용자 선호 변화**를 추론할 수 있다는 장점 때문.

(2) 어떻게 사용했는가 (방법론 구성)

프레임워크 **MLLM-SRec** 주요 단계:

1. **VQA 기반 이미지 이해 (VIU)**
   - 단순 캡셔닝 대신 질문 기반 프롬프트("이 이미지의 카테고리, 색상, 브랜드는?")를 사용해 **핵심 특징만 요약** → 시각 노이즈 제거.
2. **Item Multimodal Summary Generator (IMSG)**
   - 이미지 요약(VIU 결과) + 텍스트 설명을 LLM으로 결합 → **통합 멀티모달 표현 생성**.
3. **Temporal User Behavior Comprehension (TUBC)**
   - 시퀀스 데이터를 슬라이딩 윈도우 방식으로 처리, 시간에 따른 **사용자 선호 진화 패턴**을 추적.
4. **Supervised Fine-tuning (QLoRA 기반 경량화 SFT)**
   - 데이터셋을 instruction-input-response(Yes/No) 형식으로 변환 → 추천 태스크에 맞게 MLLM 조정.
5. **4단계 Chain-of-Thought Prompting**
   - VIU → IMSG → TUBC 결과를 단계적으로 연결해 reasoning 강화 → **최종 예측(좋아할지/싫어할지)** 수행 s41598-025-14251-1 .

---

1. 문제 상황 (Problem Situation)

- 기존 추천 시스템은 **텍스트 기반 단일 모달 데이터**(리뷰, 아이템 설명 등)에 주로 의존.
- 하지만 실제 사용자의 상호작용은 **멀티모달(이미지, 텍스트, 영상 등)** 데이터로 이루어짐.
- 기존 LLM 기반 추천 연구도 한계가 있음:
  1. **멀티모달 정보 처리 부족** – 이미지/텍스트 차이를 제대로 융합하지 못함.
  2. **시퀀스 데이터 취약** – 사용자의 선호는 시간에 따라 변하는데, 이를 잘 반영하지 못함.
  3. **멀티모달 노이즈(불필요한 시각 정보)와 의미 차이**를 처리하지 못해 추천 정확도 저하.
  4. 기존 파인튜닝 방식은 **추천 태스크에 특화된 멀티모달 최적화가 부족** s41598-025-14251-1 .

즉, **멀티모달 데이터와 동적 사용자 선호를 효과적으로 반영하지 못하는 것**이 핵심 문제입니다.

---

2. 연구 가정 (Assumptions)

연구자들이 세운 주요 가정은:

1. **MLLM(멀티모달 LLM)** 은 텍스트·이미지를 동시에 이해하고 표현을 정렬할 수 있어, 기존 문제(노이즈·모달 간 차이)를 해결할 수 있다.
2. **Instruction tuning + Chain-of-Thought prompting** 을 적용하면, MLLM의 일반 지식과 추론 능력을 **추천 태스크로 전이**할 수 있다.
3. 멀티모달 데이터를 통합하면 **데이터 희소성·cold-start 문제를 완화**하고, 더 정밀한 개인화 추천이 가능하다 s41598-025-14251-1 .

4. 실험 (Experiments)

- **데이터셋**: Amazon Review (Baby, Sports, Beauty, Toys) – 텍스트 + 이미지 활용 s41598-025-14251-1 .
- **비교 대상**:
  - 전통적 시퀀스 추천: SASRec, BERT4Rec
  - 멀티모달 추천: MMGCN, MMSR
  - LLM 기반: GPTRec, TALLRec
  - MLLM 기반: VIP5, Rec-GPT4V s41598-025-14251-1 .
- **결과** (*Table 3, p.9*):
  - 모든 데이터셋에서 **MLLM-SRec가 최고 성능**.
  - Recall, NDCG, HR에서 평균 **약 10~18% 향상** s41598-025-14251-1 .
- **Ablation 연구** (*Table 4, 5*):
  - 텍스트+이미지 조합 > 단일 모달.
  - VIU + IMSG + TUBC 조합(Mllm-SRec Full)이 가장 높은 성능 s41598-025-14251-1 .
- **Fine-tuning 전략**: QLoRA + 4-step CoT 동시 적용 시 AUC 최대 5.54% 추가 개선 s41598-025-14251-1 .

---

5. 결론 (Conclusion)

- 제안한 **MLLM-SRec**은
  - 멀티모달 데이터를 효과적으로 융합,
  - 동적 사용자 선호를 시퀀스 단위로 포착,
  - cold-start 및 데이터 희소성 문제를 완화.
- 실험적으로 **기존 SOTA 모델들을 초월**.
- 향후 연구: **추론 지연(latency)-효율성 trade-off** 개선, **확장 가능한 멀티모달 추천 아키텍처** 구축 예정 s41598-025-14251-1 .

3. 제안 방법론 (Methodology: MLLM-SRec)

프레임워크 **MLLM-SRec** 구성 요소:

1. **VQA 기반 이미지 이해 (VIU)**
   - 단순 캡셔닝 대신, 질문 기반 프롬프트("이미지에 있는 카테고리, 색상, 브랜드, 특징은?")로 핵심 시각 정보를 추출 s41598-025-14251-1 .
2. **Item Multimodal Summary Generator (IMSG)**
   - 이미지 요약(VIU 결과) + 텍스트 리뷰를 LLM으로 융합 → 통합 멀티모달 아이템 표현 생성 s41598-025-14251-1 .
3. **Temporal User Behavior Comprehension (TUBC)**
   - 슬라이딩 윈도우 방식으로 시퀀스 구성 → 시간에 따른 사용자 선호 변화를 포착.
   - LLM 기반 요약으로 사용자의 **동적 관심사**를 모델링 s41598-025-14251-1 .
4. **Supervised Fine-tuning (QLoRA 기반 PEFT)**
   - Instruction 포맷 {instruction, input, response}로 데이터 변환.
   - 예:
     - Input: "유저 프로필 + 과거 멀티모달 기록 + 후보 아이템"
     - Response: "Yes/No (선호 여부)" s41598-025-14251-1 .
5. **4단계 Chain-of-Thought Prompting**

1. 연구 문제 (Problem)

- 기존 추천 시스템(RecSys)은 대부분 텍스트 기반(단일 모달) 데이터만 활용.
- 그러나 실제 사용자는 리뷰, 이미지, 영상 등 **멀티모달 데이터**로 상호작용 → 이를 반영하지 못하면 정확성·개인화 성능이 제한됨 (arxiv.org).
- 기존 LLM 기반 추천 연구도 멀티모달 정보를 충분히 활용하지 못하고,
  - **시퀀스(순차적 행동 데이터) 처리 능력 부족.**
  - **동적 선호 변화 반영 어려움.**
  - 멀티모달 간 의미 차이와 노이즈 처리 부족 등의 한계를 보임 (arXiv).

2. 연구 가정 (Assumptions)

- **MLLM(Multimodal Large Language Model)**은 이미지·텍스트 융합 표현 학습과 세계 지식 덕분에 → 멀티모달 추천의 잠재력을 충분히 발휘할 수 있다.
- 적절한 **지시문 학습(SFT)** 과 **Chain-of-Thought(CoT)** 최적화를 통해 → LLM이 깊이 있는 사용자 추천 태스크를 수행할 수 있다 (arxiv.org).
- 멀티모달 데이터를 결합하면 **cold-start 문제**와 **데이터 희소성 문제**를 완화할 수 있다 (arXiv).