# LVLM의 멀티 뷰 이미지 처리에 관한 연구

박우제, 이인수, 성영, 심병효
서울대학교
{wjpark, islee, ysheng, bshim}@islab.snu.ac.kr

# A Study on the Multi-View Image Processing in Large Vision-Language Model (LVLM)

Park Woo Je, Lee In Su, Sheng Ying, Shim Byong Hyo
Seoul National Univ.

## 요 약

본 논문은 1인칭 시점 이미지와 3인칭 시점 이미지가 동시에 주어졌을 때 LVLM의 멀티뷰 통합과 복합적 추론 능력을 평가하고 이를 향상시키는 기법을 제안한다. 이를 위해 본 논문은 1인칭 시점 이미지와 3인칭 시점 이미지, 질문, 답변으로 구성된 100개의 쌍을 이용하여 최신 LVLM을 평가하였다. 실험 결과, 최신 LVLM에서도 60% 이하의 낮은 성능을 보였으며 이는 멀티뷰 이해 능력이 제한적임을 나타낸다. 이를 극복하기 위해 본 논문은 참조할 이미지를 결정한 후 질문에 답변하는 새로운 뷰 셀렉션 (View Selection Chain of Tought) 프롬프팅 기법을 제안한다. 제안된 기법은 기존 Chain of Thought (CoT) 기법에 비해 최대 8%의 성능 향상을 달성하였다.

## Ⅰ. 서 론

최근 LVLM은 AR, VR, 로보틱스와 같은 몰입형 환경에서 시각적 어시스턴트로 널리 활용되고 있다. 그러나 1인칭 시점 이미지는 제한된 시야와 환경 정보 부족으로 인해 다양한 질문에 대한 정확한 답변에서 한계를 보인다. 이러한 문제를 해결하기 위해 3인칭 시점 이미지를 추가하여 멀티뷰 상황에서의 시각적 이해를 강화할 필요가 있다. 이를 위해 본 논문은 3인칭 시점 이미지가 추가된 멀티 뷰 질문-답변 쌍을 구축하고, 최신 LVLM들의 성능을 평가하였다. 평가 결과, 현존하는 LVLM은 60% 이하의 낮은 성능을 보였으며 이는 LVLM이 멀티 뷰 이미지에 대한 통합적 이해에서 어려움을 겪고 있음을 시사한다. 또한, 최근 개발된 Chain of Thought (CoT) 기법을 활용하더라도 멀티뷰 이미지에 대한 추론 성능이 크게 향상되지 않는 것으로 나타났다. 이에 따라 본 논문에서는 멀티뷰 이미지를 보다 효과적으로 통합하여 LVLM의 시각적 추론 성능을 향상시키는 View Selection Chain of Thought (VSCoT) 프롬프팅 기법을 제안한다. 실험 결과, 제안된 기법은 기존 CoT 기법 대비 Gemini 2.0 Flash에서 최대 8%의 성능 향상을 보여 멀티 뷰 기반 추론의 효과를 입증하였다.

## Ⅱ. 본론

본 논문에서는 1인칭 이미지와 3인칭 이미지가 동시에 주어졌을 때 LVLM이 사용자의 질문에 적절한 답변을 생성하는 것을 목표로 한다. 이를 평가하기 위해 EgoExo4D 데이터셋을 활용하여 4가지 카테고리(동작, 색깔, 개수, 위치)에 대한 1인칭, 3인칭 이미지와 질문-답변 쌍 25개씩, 총 100개의 문제를 생성하였다. 각 카테고리의 질문은 "What am I holding?"과 "What is the person holding?"와 같이 1인칭과 3인칭 시점의 정보를 균등하게 포함하도록 설계되었다. 문제의 생성은 GPT4o-mini를 활용하였으며, 멀티뷰 이미지를 동시에 입력하여 질문과 답변, 선지를 생성한 후 전문가의 검토를 거쳐 최종적으로 100개의 4지 선다 문제를 형성하였다.

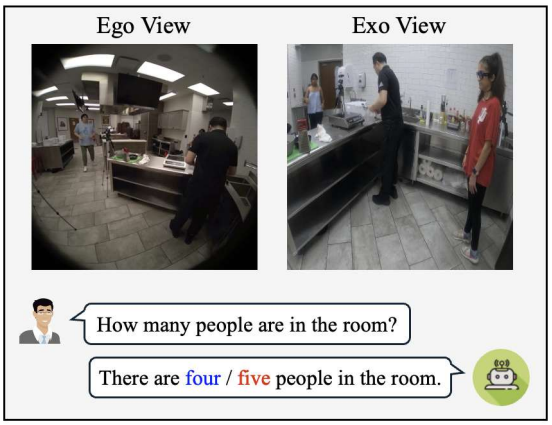


그림 1. 멀티 뷰 이미지에서 질문에 대한 정답(파란색)과 오답(빨간색)의 예시

표 1은 생성된 100개의 데이터를 기반으로 최신 LVLM의 멀티뷰 성능을 평가한 결과를 나타낸다. 여기서 일부 모델은 60%에 미치지 못하는 낮은 성능을 보였으며, 이는 다음과 같은 원인으로 분석될 수 있다: 1) 주어진 이미지에서 중요한 정보를 식별하지 못하는 경우, 2) 중복된 정보를 효과적으로 걸러내지 못하는 경우, 3) 여러 이미지를 통합하여 일관된 이해를 형성하지 못하는 경우. 그림 1의 예시는 두 이미지에 공통적으로 등장하는 객체를 고려하지 않을 때 발생하는 오류를 보여준다.

표 1. LVLM의 멀티 뷰 이미지 질문 답변에 대한 성능 평가

| LVLM | 동작 (%) | 색깔 (%) | 개수 (%) | 위치 (%) | 평균 (%) |
|---|---|---|---|---|---|
| Gemini-2.0-Flash | 52.0 | 68.0 | 40.0 | 44.0 | 51.0 |
| GPT4o-mini | 48.0 | 72.0 | 36.0 | 40.0 | 49.0 |
| InternVL3-14B | 48.0 | 68.0 | 40.0 | 36.0 | 48.0 |
| Qwen2.5-VL-7B | 52.0 | 64.0 | 40.0 | 40.0 | 49.0 |

본 논문에서는 이러한 문제를 해결하기 위해 질문의 유형에 따라 적절한 뷰를 선택하여 답변하는 VSCoT 프롬프팅 기법을 제안한다. 뷰 셀렉션은 두 개의 단계로 구성된다. 첫 번째 단계에서는 LVLM이 질문과 이미지를 바탕으로 어떤 이미지가 질문 답변에 필요한지를 결정한다. 이를 위해 3가지 관점(1인칭 이미지가 필요한 경우, 3인칭 이미지가 필요한 경우, 두

이미지 모두 필요한 경우)으로 분류하여 각 관점에 대한 이유를 모델이 형성하도록 유도하였다. 두 번째 단계에서는 이러한 관점을 바탕으로 최종 답변을 생성하도록 하여 문제를 단순화하고 성능을 향상시켰다. 표 2는 이러한 접근이 기존 CoT 기법에 비해 일관된 성능 향상을 가져오는 것을 입증하고 있다.

표 2. Gemini-2.0-Flash에서 기존 CoT 기법과 제안된 VSCoT 기법의 멀티 뷰 이미지 질문 답변에 대한 성능 평가 및 비교

| Method | 동작 (%) | 색깔 (%) | 개수 (%) | 위치 (%) | 평균 (%) |
|---|---|---|---|---|---|
| Default | 52.0 | 68.0 | 40.0 | 44.0 | 51.0 |
| CCoT | 48.0 | 72.0 | 44.0 | 40.0 | 51.0 |
| CoCoT | 56.0 | 72.0 | 44.0 | 44.0 | 54.0 |
| VSCoT (Ours) | 60.0 | 76.0 | 52.0 | 48.0 | 59.0 |

## Ⅲ. 결론

본 논문에서는 이러한 문제를 해결하기 위해 질문의 유형에 따라 적절한 뷰를 선택하여 답변하는 멀티턴 뷰 셀렉션 프롬프팅 기법을 제안한다. 뷰 셀렉션 기법은 두 개의 단계로 구성된다. 첫 번째 단계에서는 LVLM이 질문과 이미지를 바탕으로 어떤 이미지가 질문 답변에 필요한지를 결정한다. 이를 위해 3가지 관점(1인칭 이미지가 필요한 경우, 3인칭 이미지가 필요한 경우, 두 이미지 모두 필요한 경우)으로 분류하여 각 관점에 대한 이유를 모델이 형성하도록 유도하였다. 두 번째 단계에서는 이러한 관점을 바탕으로 최종 답변을 생성하도록 하여 문제를 단순화하고 성능을 향상시켰다. 표 2는 이러한 접근이 기존 CoT 기법에 비해 일관된 성능 향상을 가져오는 것을 입증하고 있다.

## ACKNOWLEDGMENT

## 참 고 문 헌


[1] Mitra, C., Huang, B., Darrell, T., & Herzig, R. (2024). Compositional chain-of-thought prompting for large multimodal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14420-14431).

[2] Zhang, D., Yang, J., Lyu, H., Jin, Z., Yao, Y., Chen, M., & Luo, J. (2024). Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. arXiv preprint arXiv:2401.02582.

[3] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[4] Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., ... & Wang, W. (2025). InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv preprint arXiv:2504.10479.

[5] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., ... & Lin, J. (2025). Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.