

# 저 사양 GPU 환경에서의 VLM-RAG 파이프라인 최적화 연구

김형오, 장문수\*

한국폴리텍대학교 성남캠퍼스 인공지능소프트웨어과,

\*한국폴리텍대학 반도체융합캠퍼스 반도체장비소프트웨어과

hokim@kopo.ac.kr, \*avecmschang@kopo.ac.kr

## A Study on VLM-RAG Pipeline Optimization in Low-Spec GPU Environments

Hyung-O Kim, Moon-su Jang\*

Dept. of AI SW, Korea Polytechnics Seongnam Campus,

\*Dept. of Semiconductor Equipment SW, Korea Polytechnics Semiconductor Convergence Campus

### 요약

대규모 시각-언어 모델인 Qwen 2.5-VL과 같은 최신 SOTA 모델은 전문 지식 분야에서 획기적인 정보 검색 능력을 제공하지만, 고비용의 고성능 GPU 자원을 요구한다. 본 논문은 이러한 하드웨어 제약을 극복하기 위한 VLM-RAG 파이프라인의 최적화 방법론을 제시한다. 실험 환경은 Google Colab의 무료 T4 GPU(VRAM 16GB) 환경에서 Qwen 2.5-VL-7B 모델을 4-bit 양자화(Quantization) 기법을 적용하여 모델 가중치를 압축함으로써 성능 저하를 최소화하면서도 모델을 저 사양 GPU 메모리에 안정적으로 로드하였다. '구석기시대 고고학 사전'이라는 전문적인 도메인 데이터에 적용된 RAG 시스템의 가장 큰 장점은 대규모 VLM의 고질적인 문제인 환각 현상(Hallucination)을 근본적으로 억제함으로써 '가로날도끼', '각추상석기' 등 특정 유물에 대한 출처 기반의 정확하고 신뢰할 수 있는 도메인 특화 지식을 제공한다는 것이다. 본 연구는 저 사양 GPU 환경에서 최신 대규모 VLM을 활용하여 전문적인 도메인 데이터에 RAG를 적용하여 성공적인 결과를 도출하였다.

### I. 서론

유물의 형태적 특징과 출토 정황, 연대 측정 데이터 등 방대한 텍스트와 이미지 정보가 유기적으로 결합된 고고학은 고도의 전문성을 요하는 복합적인 학문 영역이다. 실제로 '가로날도끼(Cleaver)'의 경우 1930년 앙리 브레이유(Henri Breuil)에 의해 용어가 정립된 이후, 티시에(Jacques Tixier)나 보르드(François Bordes)와 같은 연구자들의 관점에 따라 제작 기법과 형식 분류가 판이하게 달라지는 등 정밀한 도메인 지식이 필수적이다[1]. 하지만 범용 학습 데이터에 의존하는 일반적인 대규모 언어 모델(LLM)은 이러한 미세한 학술적 맥락을 구분하지 못해 부정확한 정보를 사실인 양 생성하는 환각 현상(Hallucination)을 빈번하게 노출한다. 이를 해결하기 위한 대안으로 외부 지식을 참조하는 검색 증강 생성(RAG) 기술이 대두되었으나[2], Qwen 2.5-VL과 같은 최신 SOTA(State Of The Art)급 시각-언어 모델(VLM)을 운용하기 위해서는 통상적으로 A100이나 H100급의 고비용 고성능 GPU 인프라가 요구된다. 이러한 물리적 하드웨어 제약은 예산이 한정된 인문학 연구 환경이나 개인 연구자들의 최신 AI 기술 접근을 원천적으로 차단하는 진입 장벽으로 작용하고 있다. 이에 본 연구는 범용적인 Google Colab의 무료 T4 GPU(16GB VRAM)급 환경에서도 최신 VLM을 원활히 구동할 수 있는 4-bit 양자화 기반의 경량화 파이프라인을 제안하고, 실제 '구석기시대 고고학 사전' 데이터를 적용하여 저 사양 환경에서의 고성능 RAG 시스템 구축 가능성과 그 효용성을 실증적으로 규명하는 것을 최종 목표로 한다.

### II. 관련연구

시각-언어 모델(Vision-Language Models, VLMs)은 대규모 텍스트-이미지 쌍을 사전 학습하여 시각적 특징과 언어적 맥락을 통합적으로 이해하는 방향으로 진화해 왔다. 특히 최신 모델인 Qwen2.5-VL은 기존 LLaVA나 CLIP 기반 모델들이 가졌던 고정 해상도 입력 방식의 한계를 극복하고, 입력 이미지의 비율에 맞춰 동적으로 토큰을 할당하는 Naive Dynamic Resolution 기술을 도입함으로써 고고학 이미지의 미세한 형태적 특징이나 텍스트 정보를 손실 없이 분석할 수 있는 강점을 지닌다[3]. 그러나 수십억 개의 파라미터를 보유한 SOTA급 VLM 모델을 제한된 GPU 메모리 환경에서 구동하는 것은 여전히 계산 비용 측면에서 도전적인 과제이다. 이에 대한 해결책으로 제시된 모델 경량화 및 양자화(Quantization) 기술은 모델의 가중치 정밀도를 낮추어 메모리 요구량을 감소시킨다. 특히 Dettmers et al.(2024)이 제안한 QLoRA 연구의 핵심인 NF4(NormalFloat 4-bit) 데이터 타입은 신경망 가중치가 0을 중심으로 하는 정규 분포를 따른다는 점에 착안하여 설계되었다[4]. 이것은 기존의 균등(Uniform) 4-bit 정수형 방식보다 양자화 오차(Quantization Error)를 획기적으로 줄이면서도 FP16(16-bit Floating Point) 대비 약 4배의 메모리 효율성을 제공한다. 본 연구는 이러한 선행 연구들의 기술적 토대 위에서 BitsAndBytes 라이브러리를 활용한 4-bit 양자화 파이프라인을 구축함으로써, 고비용 인프라 없이도 고성능 VLM의 추론 정확도를 보존하는 실용적 최적화 방법론을 채택하였다[5].

### III. 시스템 설계 및 구현

본 연구의 실험 환경은 Google Colab 플랫폼이 제공하는 NVIDIA Tesla

T4(VRAM 16GB) GPU 상에서 구축되었으며, 기반 모델로는 Qwen2.5-VL-7B-Instruct를 채택하였다. 실험 데이터로는 구체적인 유적 설명과 이미지가 포함된 고밀도 문서인 ‘한국고고학전문사전(구석기시대 편)’을 사용하였다. 고용량 PDF 처리 시 발생하는 기존 로더의 속도 저하 문제를 해결하기 위해, 텍스트 레이어를 고속으로 추출하는 PyMuPDFReader를 전처리 단계에 도입하여 데이터 로딩 시간을 단축하였다. 특히 T4 GPU의 16GB 메모리 제약을 극복하기 위해 BitsAndBytes 라이브러리를 활용한 4-bit 양자화(Quantization)를 적용하였다. NF4(NormalFloat 4-bit) 데이터 타입과 이중 양자화(Double Quantization) 설정을 통해 15GB 이상을 점유하던 모델 가중치를 약 5~6GB 수준으로 압축함으로써, 제한된 자원 내에서 RAG 검색과 컨텍스트 처리를 위한 펄스 메모리 공간을 확보하였다. RAG 파이프라인의 검색 단계에서는 다국어 처리 성능이 우수한 intfl oat/multilingual-e5-large-instruct 모델을 사용하여 고고학 전문 용어의 의미론적 벡터를 생성하였으며, 코사인 유사도(Cosine Similarity)를 기반으로 질의와 연관성이 높은 상위 3개의 문맥을 추출하여 Qwen 2.5-VL 모델이 최종 답변을 생성하도록 설계하였다..

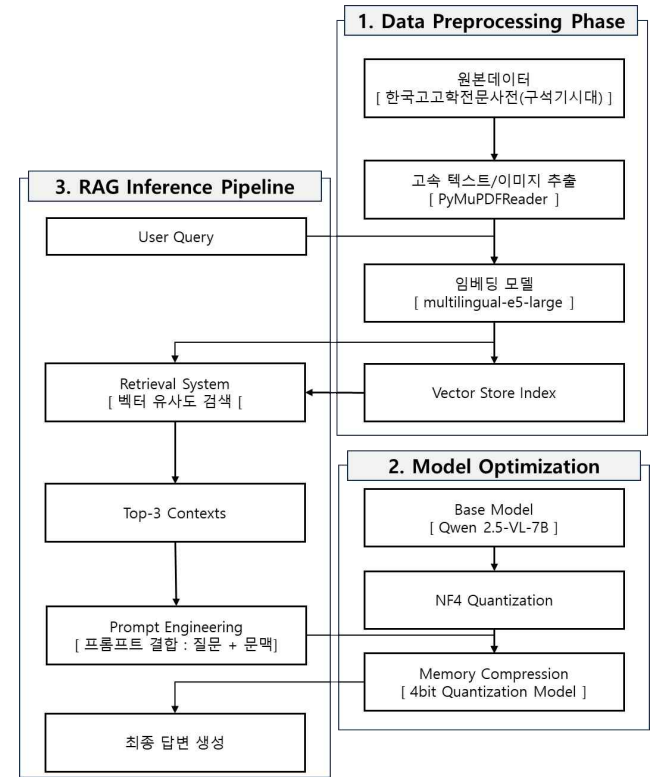


Fig 1. Flowchart of VLM-RAG System

#### IV. 실험 및 결과

본 연구에서 제안한 최적화된 RAG 파이프라인의 성능 검증은 질의응답의 정확성과 시스템 자원 효율성 두 가지 측면에서 수행되었다. 먼저 정성적 평가를 위해 수행된 “가로날도끼의 특징” 질의 실험에서, 모델은 해당 유물이 전기 구석기시대 아슐리안 문화 단계의 산물임을 식별하고, 텍시에(Jacques Tixier)의 기술적 분류 기준인 ‘잔손질이 없는 자연 상태의 날’이라는 핵심 정보를 정확하게 인출(Retrieval)하였다. 나아가 범용 모델이 놓치기 쉬운 칼람보 폭포 유적의 실증적 사용 흔적과 브레이유, 보프레이 등 주요 연구자들의 이론적 배경까지 문맥 정보에 기반하여 생성함으로써 도메인 특화 지식의 전달력이 입증되었다. 시스템 성능 평가에서는 4-bit 양자화(Quantization) 기술을 통해 모델의 GPU VRAM 점유율을 약 10.5GB로 제어하여, 16GB 메모리를 가진 Colab T4 환경에서의 안정

적인 구동을 실현하였다. 특히 데이터 전처리 단계에서 PyMuPDFReader를 적용하여 기존 방식 대비 인덱싱 대기 시간을 15분에서 수 초 단위로 단축하는 속도 개선을 달성하였으며, 평균 15~20초의 추론 시간은 실시간 대화에는 다소 한계가 있으나 고심도 연구 분석용으로는 유효한 수준임을 확인하였다.

RAG 응답 결과	고고학 사전 PDF 원문
<p>=== 질의응답 테스트 ===</p> <p>질문: 가로날도끼의 특징과 이것이 처음 발견된 곳에 대해 설명해줘.</p> <p>답변:</p> <p>가로날도끼는 전기 구석기시대 아슐리안 문화 단계에서 많이 제작된 연모입니다. 이 도구는 끝에 날이 있으며, 사용 중에 으러진 자국을 제외하고 날은 잔손질하지 않은 자연 상태라고 정의됩니다. 격지의 밀면과 격지 뒷면이 만나서 날을 형성하기 때문에 격지를 이용해서 만든 것만 가로날도끼로 분류됩니다.</p> <p>가로날도끼는 아프리카를 비롯하여 유럽의 남서부 및 아시아 일부 지역에서 발견되며, 이 용어는 1930년에 Henri Breuil이 아프리카에서 출토된 석기를 설명하기 위해 처음 사용하였습니다. 그러나 가로날도끼에 대한 정의와 분류는 연구자마다 다릅니다. Jacques Tixier는 가로날도끼를 격지의 밀면과 뒷면이 만나서 날을 형성하는 연모로 정의하며, François Bordes는 가로날도끼를 양면석기의 하나로 분류하였습니다.</p> <p>또한, 아프리카의 칼람보 폭포(Kalambo Falls) 유적에서 출토된 가로날도끼의 날 부분에서 사용된 흔적과 망가진 흔적이 확인되어, 이 도구가 동물의 뼈와 고기를 가르는 데 사용된 것으로 추정됩니다.</p>	<p><b>가로날도끼</b> Cleaver</p> <p>가로날도끼는 전기 구석기시대 아슐리안 문화 단계에서 많이 제작된 연모이다. 아프리카를 비롯하여 유럽의 남서부 및 아시아 일부 지역에서 발견된다. 이 용어는 브레이유(Henri Breuil)가 1930년에 아프리카에서 출토된 석기를 설명하기 위해 처음 사용하였다. 가로날도끼는 연구자에 따라서 제작 기법과 형식 분류에 많은 차이가 있다. 텍시에(Jacques Tixier)는 가로날도끼는 끝에 날이 있는 연모이며, 사용 중에 으러진 자국을 제외하고 날은 잔손질하지 않은 자연 상태라고 정의하였다. 이 연모는 격지의 밀면과 격지 뒷면이 만나서 날을 형성하기 때문에 격지를 이용해서 만든 것만 가로날도끼로 분류하였다. 아프리카에서 출토된 유물을 기술적인 특징을 기준으로 하여 6가지로 세분화했다. 이에 반해 보르드(François Bordes)는 가로날도끼를 양면석기의 하나로 분류하였는데, 대체로 두툼고 밀면과 마주보는 곳에 가로로 된 날이 있다고 하였다. 보프레이(Raymond Vaufray)도 가로날도끼를 가로날을 갖는 양면석기로 분류하였다. 이 도구는 동물의 뼈와 고기를 가르는 데 사용된 것으로 추정되는데, 실제로 아프리카의 칼람보 폭포(Kalambo Falls)유적에서 출토된 가로날도끼의 날 부분에서 사용된 흔적과 망가진 흔적이 확인되었다. (공수진)</p>

Fig 2. (Left) VLM-RAG Inference result (right) PDF Source

#### V. 결론

본 연구는 하드웨어 제약을 극복하기 위한 VLM-RAG 파이프라인의 최적화 방법론을 제시하였다. Google Colab의 T4 GPU(VRAM 16GB) 환경에서 Qwen 2.5-VL-7B 모델을 4-bit 양자화(Quantization) 기법을 적용하여 모델 가중치를 압축함으로써 성능 저하를 최소화하면서도 저사양 GPU에 안정적으로 로드하였다.

‘구석기시대 고고학 사전’이라는 전문적인 도메인 데이터를 PyMuPDFReader 기반의 고속 데이터 전처리 RAG파이프라인 시스템을 구현하여 하드웨어 자원의 제약을 극복하고 효율적인 결과를 도출하였다.

#### 참 고 문 헌

[1] Lee, Y., et al. (2012). Dictionary of Korean Archaeology: The Paleolithic Age. Seoul: National Research Institute of Cultural Heritage.

[2] Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. International Conference on Machine Learning (ICML), 15696-15707.

[3] Wang, Peng et al. “Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution.” ArXiv abs/2409.12191 (2024): n. pag.

[4] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv, abs/2305.14314..

[5] Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. ArXiv, abs/2208.07339..

