# 시각-언어 모델 기반 문화유산 이미지 설명의 한-일 이중언어 번역

박세현, 노동원, 장한얼, 김태훈, 박천음*

국립한밭대학교

20242174@edu.hanbat.ac.kr, nohdw@edu.hanbat.ac.kr, {hejang, thkim, parkce}@hanbat.ac.kr

# Korean-Japanese Bilingual Translation of Cultural Heritage Image Descriptions with VLM

Sehyun Park, Dongwon Noh, Haneol Jang, Taehoon Kim, Cheoneum Park*

Hanbat National University

요 약

본 논문은 문화유산 설명문 번역에서 텍스트 중심 접근의 한계를 보완하기 위한 시각 정보 활용 가능성을 분석한다. 이를 위해 일본 문화유산 데이터베이스 ColBase를 기반으로 이미지-텍스트 쌍으로 구성된 일본어-한국어 양방향 번역 데이터셋을 구축하고, 텍스트 단독 입력과 이미지 결합 입력 조건에서 멀티모달 번역 실험을 수행하였다. 실험 결과, 시각 정보는 일부 모델과 번역 방향에서 BLEURT 및 COMET과 같은 의미 기반 평가 지표의 향상을 보이며 번역 품질 개선에 기여하였으나, 모든 설정에서 일관된 성능 향상으로 이어지지는 않았다. 이러한 결과는 문화유산 번역에서 시각 정보가 이점을 가지는 동시에, 번역 대상과 모델 특성에 따라 선별적으로 활용될 필요가 있음을 보였다.

## I. 서론

전 세계적으로 산업 전반의 디지털 전환이 가속됨에 따라, 문화유산 분야 역시 단순한 물리적 보존을 넘어 디지털 아카이빙을 기반으로 한 공유화 확산의 단계로 빠르게 전환되고 있다. 고해상도 이미지와 상세한 설명문이 결합된 디지털 문화유산 데이터는 시공간적 제약을 완화함으로써 국가 간 문화 접근성을 확대하고, 국제적 문화 교류를 촉진하는 핵심 인프라로 기능한다. 문화재 정보를 상호 언어로 정확하게 전달할 수 있는 고품질 번역의 중요성이 지속적으로 부각되고 있으며, 문화재 설명문은 유물 고유의 가치와 역사적, 문화적 맥락을 함께 전달해야 하므로 일반 도메인 번역과는 구별되는 전문성이 요구된다 [1].

기존 연구에 따르면, 문화재 번역에서 이미지를 함께 활용하는 멀티모델 접근은 텍스트 정보만으로 포착하기 어려운 유물의 형태적·맥락적 정보를 보완함으로써 보다 적절한 번역을 가능하게 할 잠재력을 지닌다 [2, 3]. 또한 시각 정보는 유물의 제작 시기와 문화적 맥락을 반영하는 단서로 활용될 수 있으며, 이를 통해 한자 고유명사의 역사적 의미 해석을 보완할 가능성이 제시되어 왔다.

본 연구는 일본의 대표적인 문화유산 데이터베이스인 ColBase를 기반으로 한국어와 일본어 양방향 번역 과정에서 시각 정보가 번역 정확도 향상 분석을 수행한다. 이를 위해 이미지-텍스트 쌍으로 구성된 문화재 번역용 데이터셋을 구축하고, 여러 멀티모달 대규모 언어 모델을 활용한 비교 실험을 수행한다. 본 연구의 주요 기여는 다음과 같다. 첫째, 일본 문화재 설명문 번역을 위한 이미지-텍스트 병렬 골드 데이터셋을 구축한다. 둘째, 시각 정보의 유무에 따른 번역 성능 차이를 자동 평가 지표를 통해 정량적으로 분석한다.

*Corresponding author

## II. 제안 방법

본 논문은 시각 정보가 일본 문화재 설명문의 번역 품질에 미치는 영향을 분석하기 위해, 고품질 평가 데이터셋을 구축하고 이를 기반으로 멀티모달 번역 실험을 수행한다. 기존 텍스트 위주의 문화재 번역 연구와 달리, 본 연구는 이미지 정보를 명시적으로 번역 입력으로 포함하여 시각 정보의 기여도를 체계적으로 분석한다.

### II.I. 데이터 수집 및 선별

본 연구에 사용된 원본 데이터는 일본 국립문화재기구(National Institutes for Cultural Heritage)가 운영하는 통합 문화유산 데이터베이스 ColBase [4]로부터 수집하였다. ColBase는 일본 전역의 국공립 박물관 및 연구기관이 소장한 문화재의 이미지와 메타데이터를 체계적으로 제공하는 디지털 문화유산 아카이브로, 회화, 조각, 공예, 서적 등 다양한 유형의 문화재를 포괄한다.

데이터 수집 단계에서는 각 문화재의 명칭, 이미지, 일본어 설명문을 포함하여 총 160,000개의 원천 데이터를 확보한다. 이후 문화적, 역사적 맥락을 반영하기 위해 연대별 분포를 기준으로 구분하고, 일본 사회의 변화가 활발히 반영된 기원전 300년부터 1900년대 구간의 데이터를 선별한다. 특히 역사적 사건, 사회 제도, 종교 및 관습과 관련된 서술을 포함한 문서를 우선적으로 선정하여 8,000개의 데이터 샘플을 구축한다. 이미지가 누락된 경우, 설명문의 길이가 번역 평가에 적합하지 않을 정도로 길거나 짧은 경우를 제외하여 최종적으로 370개의 고품질 데이터를 선별한다.

### II.II. 평가 데이터 구축

선별된 370개의 데이터에 대해 3단계 절차를 통해 일본어, 한국어 양방향 번역을 위한 골드 데이터셋을 구축한다. 먼저, ColBase에서 수집된 원문에 포함된 HTML 태그, 불필요한 공백을 제거하

표 1: JP–KR / KR–JP Translation Performance Comparison

| Target Language | Model | Type | BLEU | BLEURT | COMET | ROUGE-1 |
|---|---|---|---|---|---|---|
| JP to KR | Qwen3-VL-8B-Instruct | Text-Only | **14.43** | 49.22 | 79.41 | **51.31** |
| | | Text+Image | 13.61 | **49.57** | **79.49** | 49.91 |
| | gemma-3n-E4B-it | Text-Only | 14.25 | 46.40 | 75.16 | 47.40 |
| | | Text+Image | **15.71** | **49.22** | **77.42** | **48.37** |
| KR to JP | Qwen3-VL-8B-Instruct | Text-Only | 1.44 | 58.52 | 84.20 | 24.33 |
| | | Text+Image | **3.47** | **59.73** | **84.48** | **24.65** |
| | gemma-3n-E4B-it | Text-Only | 0.36 | 53.31 | **83.50** | **23.76** |
| | | Text+Image | **1.59** | **56.14** | 83.16 | 23.58 |

여 텍스트의 가독성과 처리 효율성을 확보한 후, 대규모 언어 모델을 활용한 1차 번역을 수행한다. OpenAI의 GPT-4o mini 모델을 사용하여 일본어 원문을 한국어로 번역하고, 이 과정에서 문화재 번역에 적합한 프롬프트를 설계하여 전문 용어의 정확성과 문맥의 충실도를 확보한다. 마지막으로, 일본어와 한국어 검수자가 1차 번역 결과를 원문과 대조하며 검수를 진행한다. 검수 과정에서 의미 누락 또는 왜곡, 과도한 의역 등의 오류 유형을 중점적으로 검토하고 수정한다. 이러한 절차를 통해 구축된 학습 데이터셋은 높은 번역 품질과 용어 일관성을 확보한다.

## III. 실험

시각 정보의 유무에 따른 번역 성능 차이를 검증하기 위해, 본 연구는 텍스트 정보만을 입력으로 사용하는 경우(Text-Only)와 동일한 텍스트에 문화재 이미지를 함께 제공하는 경우(Text + Image)로 실험을 수행한다. 실험에 사용된 모델은 Qwen3-VL-8B-Instruct, gemma-3n-E4B-it이며, 평가지표는 BLEU, BLEURT, COMET-22-DA(이하 COMET), ROUGE-1 을 사용한다. 모든 실험에서 입력 텍스트와 프롬프트 구조는 동일하게 유지하여 이미지 입력 여부에 따른 성능 차이를 분석하는 방식으로 수행한다.

### III.I. 실험 결과

표 1는 시각 정보의 유무에 따른 일본어–한국어(JP→KR) 및 한국어–일본어(KR→JP) 번역 성능을 비교한 결과를 제시한다.

전반적으로 시각 정보를 함께 제공한 경우(Text + Image)는 일부 설정에서 텍스트만을 입력으로 사용한 경우(Text-Only)에 비해 전반적으로 성능 향상을 보였으나, 이러한 경향은 모든 모델과 번역 방향에서 일관되게 나타나지 않았다.

JP→KR 번역의 경우, gemma-3n-E4B-it 모델은 Text + Image 조건에서 BLEU, BLEURT, COMET, ROUGE-1 점수가 모두 상승하며 시각 정보의 긍정적인 효과를 보였다. 반면 Qwen3-VL-8B-Instruct 모델에서는 BLEURT와 COMET 점수가 소폭 향상되었으나, BLEU와 ROUGE-1 점수는 오히려 감소하는 경향이 관찰되었다. 이는 이미지 입력이 의미적 판단에는 기여했으나, 형태적 일치도를 중시하는 지표에서는 오히려 불리하게 작용했을 가능성을 시사한다.

KR→JP 번역에서는 두 모델 모두 Text + Image 조건에서 BLEU와 BLEURT 점수가 일관되게 개선되었으며, 특히 Qwen3-VL-8B-Instruct 모델에서 BLEU 점수가 1.44에서 3.47로 크게 상승하였다. 이는 시각 정보가 의미 보완을 통해 일본어 번역에서 보다 적절한 표현 선택을 유도했음을 시사한다. 반면 gemma-3n-E4B-it 모델에서는 ROUGE-1 점수가 소폭 하락하였는데, 이는 이미지 입력이 어휘 수준의 중복을 증가시키기보다는, 문맥적 의미 보존과 적절한 표현을 유도함으로써 번역의 의미적 충실도를 향상시켰기 때문으로 해석할 수 있다.

이러한 결과는 시각 정보가 번역 품질 향상에 기여할 수 있으나, 항상 모든 평가 지표에서 성능 향상으로 이어지지는 않음을 시사한다. 특히 텍스트 정보만으로도 충분한 정보가 제공되는 경우, 이미지 입력이 추가적인 잡음으로 작용할 가능성이 있으며, 본 연구에서 사용된 비교적 소규모 파라미터의 모델 특성을 고려할 때 시각 정보와 텍스트 정보를 정교하게 통합하는 데 구조적 한계가 존재했을 가능성도 배제할 수 없다.

## IV. 결론

본 연구는 문화유산 설명문 번역이라는 도메인 특화 환경에서 시각 정보가 기계 번역 품질에 미치는 영향을 분석하였다. 이를 위해 일본 문화유산 데이터베이스 ColBase를 기반으로 이미지–텍스트 쌍으로 구성된 일본어–한국어 양방향 번역 데이터셋을 구축하고, 텍스트 단독 입력과 이미지 결합 입력 조건을 비교하는 멀티모달 번역 실험을 수행하였다. 실험 결과, 시각 정보를 함께 제공한 경우(Text + Image)는 일부 모델 및 번역 방향에서 BLEU, BLEURT, COMET, ROUGE-1 등의 지표에서 성능 향상을 보였으나, 이러한 효과는 모든 설정에서 일관되게 나타나지는 않았다. 특히 텍스트 정보만으로도 충분한 의미 해석이 가능한 경우에는 이미지 입력이 성능 향상으로 이어지지 않는 경우도 확인되었다. 이러한 결과는 문화유산 설명문 번역에서 시각 정보가 잠재적 이점을 지니는 동시에, 모델 구조와 입력 조건에 따라 그 효과가 달라질 수 있음을 시사한다. 본 연구에서 사용된 비교적 소규모 파라미터의 멀티모달 모델 특성을 고려할 때, 시각 정보와 텍스트 정보를 정교하게 통합하는 데 한계가 존재했을 가능성도 배제할 수 없다.

향후 연구로는 본 연구를 다양한 문화권으로 확장하고, 다개국 문화유산 해석 및 교육 환경에서의 실제 응용을 반영한 데이터 확장과 함께, 동아시아 문화유산 간 연관성을 분석하는 멀티모달 연구로 범위를 확장하고자 한다.

## 사 사 문 구

본 과제(결과물)는 2025년도 교육부 및 대전광역시의 재원으로 대전 RISE센터의 지원을 받아 수행된 지역혁신중심 결과입니다. (2025-RISE-06-002)

## 참 고 문 헌

[1] J. Li, D. Ataman, and R. Sennrich, "Vision matters when it should: Sanity checking multimodal machine translation models," *arXiv preprint*, 2021.

[2] A. Hatami, M. Arcan, and P. Buitelaar, "Leveraging visual scene graph to enhance translation quality in multimodal machine translation," in *Proceedings of Machine Translation Summit XX: Volume 1* (P. Bouillon, J. Gerlach, S. Girletti, L. Volkart, R. Rubino, R. Sennrich, A. C. Farinha, M. Gaido, J. Daems, D. Kenny, H. Moniz, and S. Szoc, eds.), (Geneva, Switzerland), pp. 353–364, European Association for Machine Translation, Jun 2025.

[3] E. Villa-Cueva, S. Bolatzhanova, D. Turmakhan, K. Elzeky, H. B. Ademtew, A. F. Aji, V. Araujo, I. A. Azime, J. Baek, F. Belcavello, F. Cristobal, J. C. B. Cruz, M. Dabre, R. Dabre, T. Ehsan, N. A. Etori, F. Farooqui, J. Geng, G. Ivetta, T. Jayakumar, S. Jeong, Z. W. Lim, A. Mandal, S. Martinelli, M. M. Mihaylov, D. Orel, A. Pramanick, S. Purkayastha, I. Salazar, H. Song, T. T. Torrent, D. D. Yadeta, I. Hamed, A. L. Tonja, and T. Solorio, "Cammt: Benchmarking culturally aware multimodal machine translation," *arXiv preprint*, 2025.

[4] K. KAWAI, C. ODA, and N. TSURUGA, "Usage status of cultural property photographs on "colbase"," *Dejitaru Akaibu Gakkaishi*, vol. 10, no. 1, pp. e1–e10, 2025.