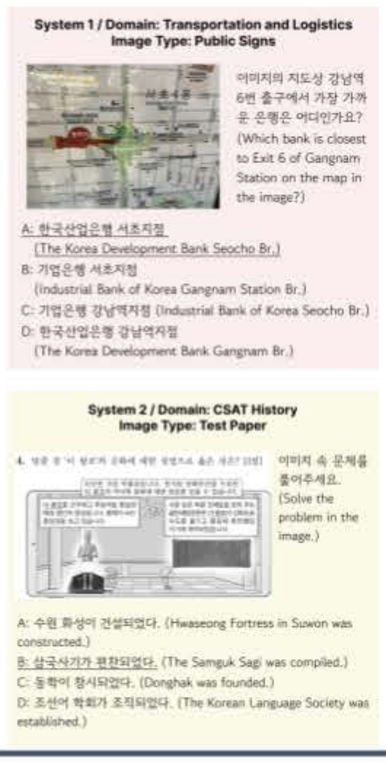
KRETA: A Benchmark for Korean Reading and Reasoning in Text-Rich VQA Attuned to Diverse Visual Contexts

Taebaek Hwang* · Minseo Kim* · Gisang Lee · Seonuk Kim · Hyunjun Eun

1 INTRODUCTION

1.1 Background

- **Text embedded within images** is crucial for conveying information across various real-world domains.
- Extensive VQA research focuses on **text-rich images (e.g., documents, scene text, digital interfaces)** to advance VLM capabilities.
- Recent benchmarks emphasize **higher-order reasoning** over textual content, moving beyond basic recognition.



1.2 Problem Statement

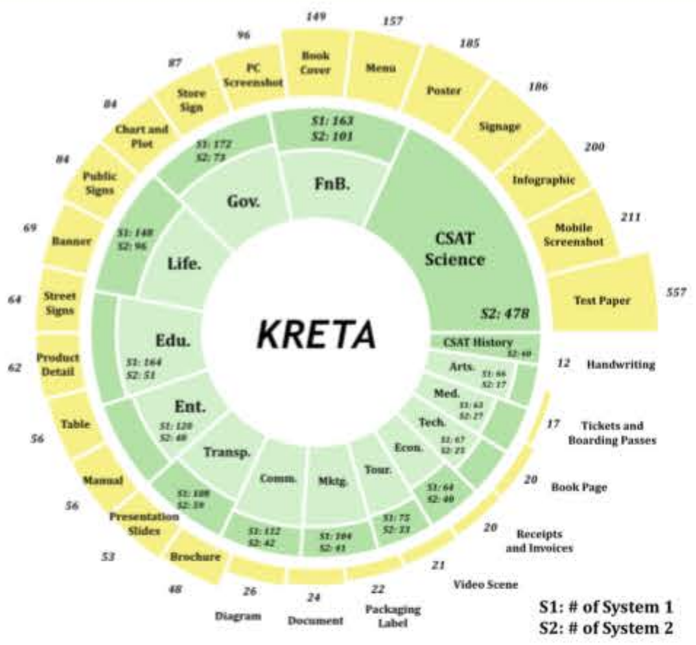
- Despite the existence of text-rich VQA benchmarks in English, Korean resources remain **scarce**, as current datasets are mostly translated or **too small in scale**.

Benchmark	Image Source	Samples	Text-Centric Reasoning	Forms	Image Type
K-MMB (Ju et al., 2024)	En	4,329	-	MC	General
K-SEED (Ju et al., 2024)	En	2,971	-	MC	General
K-MMSTAR (Ju et al., 2024)	En	1,500	-	MC	General
K-LLaVA-W (Ju et al., 2024)	En	60	-	Open	General
K-Viscuit (Baek et al., 2024)	Ko	657	-	MC	General
K-DTCBench (Ju et al., 2024)	Ko	240	✓	MC	Document
MTVQA-ko (Tang et al., 2024b)	Ko	558	-	Short	Multi-text
KOFFVQA (Kim and Jung, 2025)	Ko	275	✓	Open	Multi-text
KRETA (Ours)	Ko	2,577	✓	MC	Multi-text

2 KRETA Benchmark

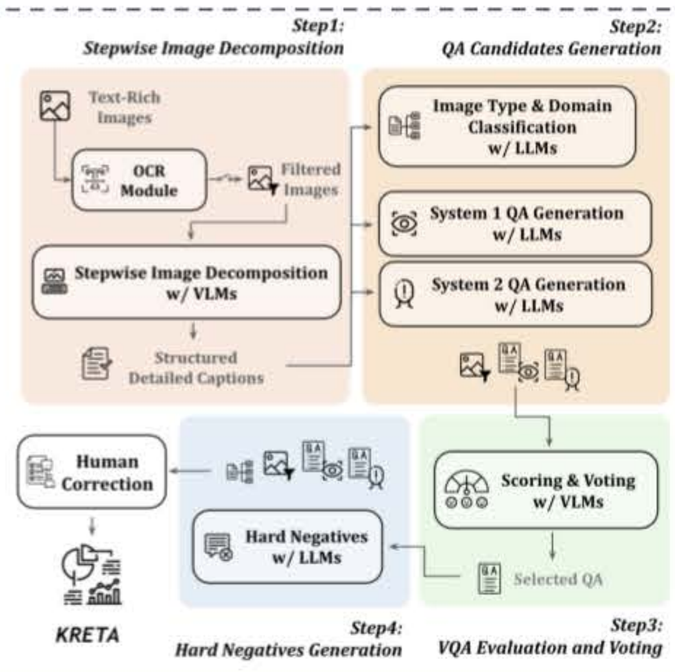
2.1 Overview

- **System 1 (Basic Understanding):** Relies on intuitive and automatic recognition. Requires direct text extraction and straightforward interpretation (1,426 QA pairs).
- **System 2 (Advanced Reasoning):** Demands complex "slow thinking," such as contextual understanding, multi-step decision-making, numerical reasoning, and integration of external knowledge (1,151 pairs)
- **15 Domains & 26 Image Types**



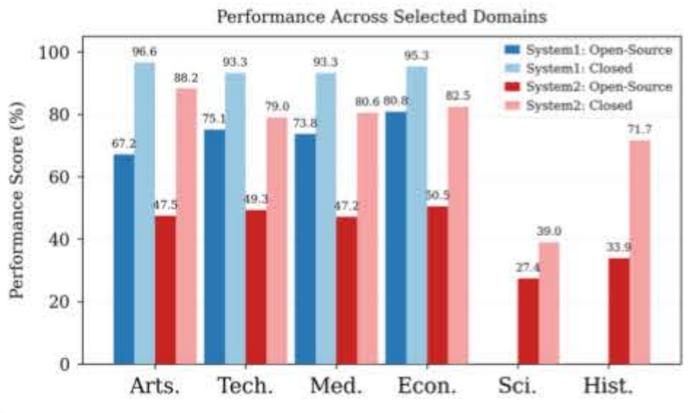
2.2 Semi-Automated VQA Generation Pipeline

- **Step 1: Stepwise Image Decomposition:** Generate structured captions from text-rich images using multiple VLMs; analyze scene/layout, link text–visual relations, then extract & structure all text.
- **Step 2: QA Candidates Generation:** From captions, create QA candidates with a dual-level setup (S1 basic recognition, S2 advanced reasoning); also tag 15 domains and 26 image types.
- **Step 3: VQA Evaluation and Voting:** Score with multi-VLM evaluation on 7 metrics; aggregate and vote to select the best QA per image.
- **Step 4: Hard Negatives & Human Refinement:** Generate hard negatives (plausible but wrong); finalize with human verification for context and reasoning validity.



3 Evaluation

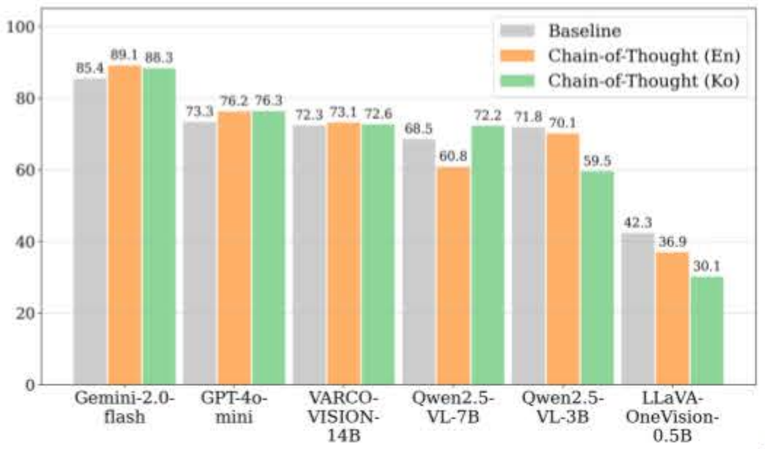
- System 1 is strong overall, but System 2 drops sharply (especially for open-source)
- Closed >> Open on System 2
- Open models vary by domain, scene-text & complex docs remain hardest
- Scaling helps but not always (Qwen2.5-VL 3B > 7B)1



Model	Size	Overall (2,577)	System 1 (1,426)	System 2 (1,151)
Closed				
GPT-4o (OpenAI, 2024a)	-	84.6	95.9	**70.5**
GPT-4o-mini (OpenAI, 2024a)	-	73.3	88.7	54.1
Gemini-2.0-flash (Google DeepMind, 2025)	-	**85.4**	**98.0**	69.8
Claude-3.5-Sonnet (Anthropic, 2024)	-	80.5	93.4	64.5
Open-source				
LLaVA-OneVision (Li et al., 2024)	0.5B	42.3	49.6	33.3
Deepseek-VL2-tiny (Wu et al., 2024)	1B	48.8	60.8	34.0
Deepseek-VL2-small (Wu et al., 2024)	2.8B	53.3	67.3	36.1
Qwen2.5-VL (Wang et al., 2024)	3B	**71.8**	**94.2**	43.9
Ovis1.6-Llama3.2 (Lu et al., 2024)	3B	52.2	62.8	39.1
InternVL2.5 (Chen et al., 2024b)	4B	70.7	90.7	**45.9**
Phi-3.5-Vision (Abdin et al., 2024)	4.2B	42.6	52.2	30.8
LLaVA-OneVision (Li et al., 2024)	7B	54.0	65.1	40.1
Qwen2.5-VL (Wang et al., 2024)	7B	68.5	**94.5**	36.1
InternVL2.5 (Chen et al., 2024b)	8B	70.8	89.8	47.3
MiniCPM-V-2.6 (Yao et al., 2024)	8B	41.0	50.4	29.4
MiniCPM-o-2.6 (Yao et al., 2024)	8B	64.3	84.1	39.9
Ovis1.6-Gemma2 (Lu et al., 2024)	9B	58.4	68.9	45.4
VARCO-VISION (Ju et al., 2024)	14B	**72.3**	90.9	**49.3**

4 Discussion - CoT

- CoT consistently benefits **closed** & **large open models** (En/Ko)
- **Smaller open models** show little or negative gain, likely from overload under long reasoning prompts.



GitHub https://github.com/tabtoyou/KRETA Hugging Face https://huggingface.co/datasets/tabtoyou/KRETA