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

Anonymous EMNLP submission

Abstract

Understanding and reasoning over text within visual contexts poses a significant challenge for Vision-Language Models (VLMs), given the complexity and diversity of real-world scenarios. To address this challenge, text-rich Visual Question Answering (VQA) datasets and evaluation benchmarks have emerged for high-resource languages like English. However, a critical gap remains: the lack of comprehensive, high-quality benchmarks for low-resource languages such as Korean, which hinders reliable model development and comparison. To bridge this gap, we introduce **KRETA**, a benchmark for **K**orean **R**eading and r**E**asoning in **T**ext-rich VQA **A**ttuned to diverse visual contexts. KRETA facilitates an in-depth evaluation of both visual text understanding and reasoning capabilities, while also supporting a multi-faceted assessment across 15 domains and 26 image types. Additionally, we introduce a semi-automated VQA generation pipeline specifically optimized for text-rich settings, leveraging refined stepwise image decomposition and a rigorous seven-metric evaluation protocol to ensure data quality. We hope that our generation pipeline will be adaptable to other languages, accelerating multilingual VLM research. The code and dataset for KRETA are available at `anonymous.4open.science`.

1 Introduction

In real-world scenarios, text within images plays a crucial role in conveying information across various domains. Thus, extensive research in VQA has focused on text-rich images, such as documents (Mathew et al., 2021; Masry et al., 2022), scene text (Singh et al., 2019; Mishra et al., 2019), and digital interfaces (Hsiao et al., 2022), driving advances in Vision-Language Models (VLMs) (Liu et al., 2023a; Wang et al., 2024; Zhang et al., 2024b) designed to handle these diverse visual contexts. Recently, the field has progressed beyond basic text recognition, with new benchmarks (Yue et al., 2024c; Hao et al., 2025) emphasizing higher-order reasoning over textual content within images. Addressing these challenges necessitates tightly integrated cross-modal understanding, leveraging domain knowledge and multi-step reasoning that cannot be achieved by treating visual and linguistic elements in isolation.

However, low-resource languages including Korean lack benchmark suites even for basic text recognition, much less reasoning, impeding comprehensive evaluation and hindering model development across diverse domains (e.g., commerce, education) and image types (e.g., street signs, charts). Although recent multilingual VQA benchmarks (Tang et al., 2024b; Sun et al., 2024) have begun to address this disparity, they often struggle to provide sufficient coverage and depth for all languages. Existing Korean VQA datasets (Ju et al., 2024; Kim and Jung, 2025) often rely on translated English questions and non-Korean images, or are limited in scale (e.g., fewer than 650 samples).

To fill the underexplored evaluation gap for Korean text-rich VQA, we propose **KRETA**, a benchmark for **K**orean **R**eading and r**E**asoning in **T**ext-rich VQA **A**ttuned to diverse visual contexts. Specifically, Figure 1 (a) shows how KRETA is built upon a wide range of real-world Korean imagery, which we systematically categorized into 15 domains by referring to the Korean Standard Industrial Classification (KSIC) (Korea Statistics, 2024) and 26 image types widely used in prior works (Yue et al., 2024a; Tang et al., 2024b). Furthermore, we carefully design a dual-level reasoning framework inspired by the concepts of *System 1* and *System 2* (Kahneman, 2011): *System 1* assesses basic text recognition, while *System 2* evaluates advanced capabilities such as domain-specific knowledge understanding, multi-step reasoning, and visual-based mathematical reasoning. KRETA comprises 2,577 samples, including 1,426 *System 1* QA pairs and

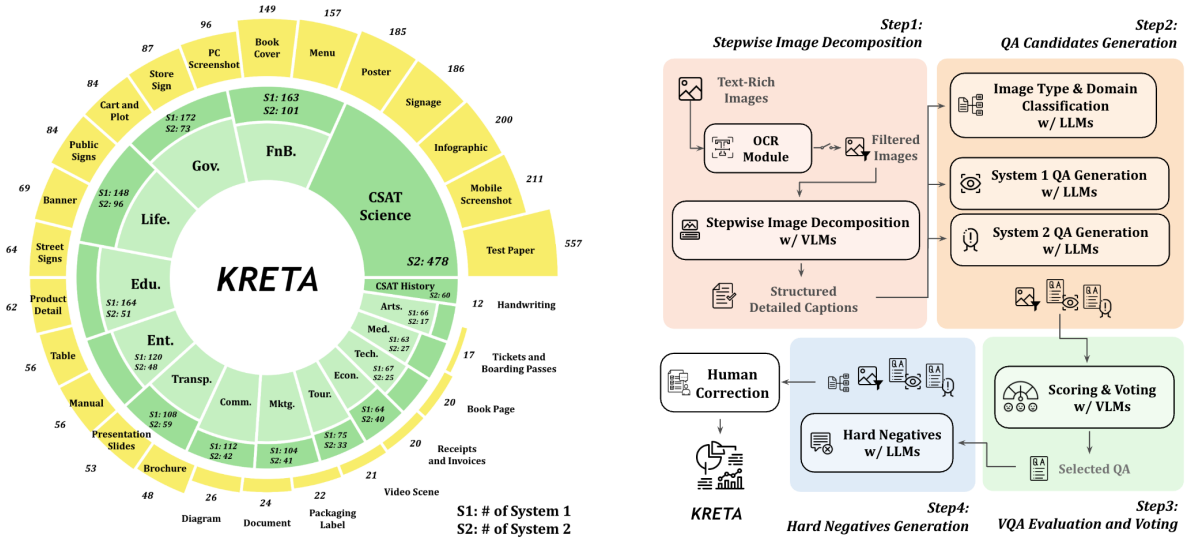

Figure 1: (a) Distribution of samples across 15 domains (inner ring) and 26 image types (outer ring). Dark green and light green segments in the inner ring represent the number of samples associated with System 2 and System 1, respectively. See Subsection 3.1 for domain abbreviations. (b) The semi-automated VQA generation pipeline.

1,151 *System 2* QA pairs, and is, to the best of our knowledge, among the largest Korean text-rich VQA datasets currently available.

To ensure scalability and quality, we design a semi-automated VQA generation pipeline, as illustrated in Figure 1 (b). Unlike prior approaches (Chen et al., 2024a), our method is specifically tailored for text-rich settings, centering on a refined, stepwise and multi-model decomposition that merges multiple VLM outputs to create high-quality structured captions for each image. This process is critical not only for capturing both textual and visual context, but also for minimizing hallucinations. Using these captions, we generate and evaluate QA candidates, synthesize hard negatives, and conduct final human refinement to ensure benchmark fidelity. We also release all prompts for question generation, as well as our seven evaluation metrics specifically designed for text-rich VQA, to support transparent adaptation and reproducibility.

Finally, our empirical analysis leveraging KRETA reveals that while VLMs demonstrate proficiency in basic Korean text recognition (System 1), a significant bottleneck remains for higher-order tasks requiring multi-step reasoning (System 2), particularly in open-source models. These models notably struggle with domain-specific knowledge and complex layouts, showing pronounced difficulty in areas like CSAT History and Marketing, as well as with image types such as banners and store signs. This underscores the need for targeted training on data encompassing Korean cultural and domain-specific knowledge, complex real-world layouts, and multi-step reasoning tasks. Our key contributions are threefold:

1. **An in-depth and multi-faceted evaluation framework:** We adopt a dual-level reasoning framework, *System 1* for basic understanding and *System 2* for advanced reasoning, to provide an in-depth evaluation of VLM performance on text-rich images. Additionally, we adopt a multifaceted classification framework for images based on *domain* and *image type* to facilitate task-specific usage and evaluation in real-world industrial applications.

2. **A semi-automated VQA generation pipeline:** We present a systematic and scalable pipeline optimized for text-rich VQA, featuring refined stepwise image decomposition and a seven-metric evaluation protocol to ensure data quality. To support adaptation to other low-resource languages, we release not only the dataset but also all prompts and code.

3. **A comprehensive text-rich VQA benchmark for Korean:** By integrating the above approaches, KRETA offers the first large-scale, high-quality benchmark to assess both basic and advanced reasoning of VLMs on real-world, text-rich Korean images spanning diverse domains and image types.

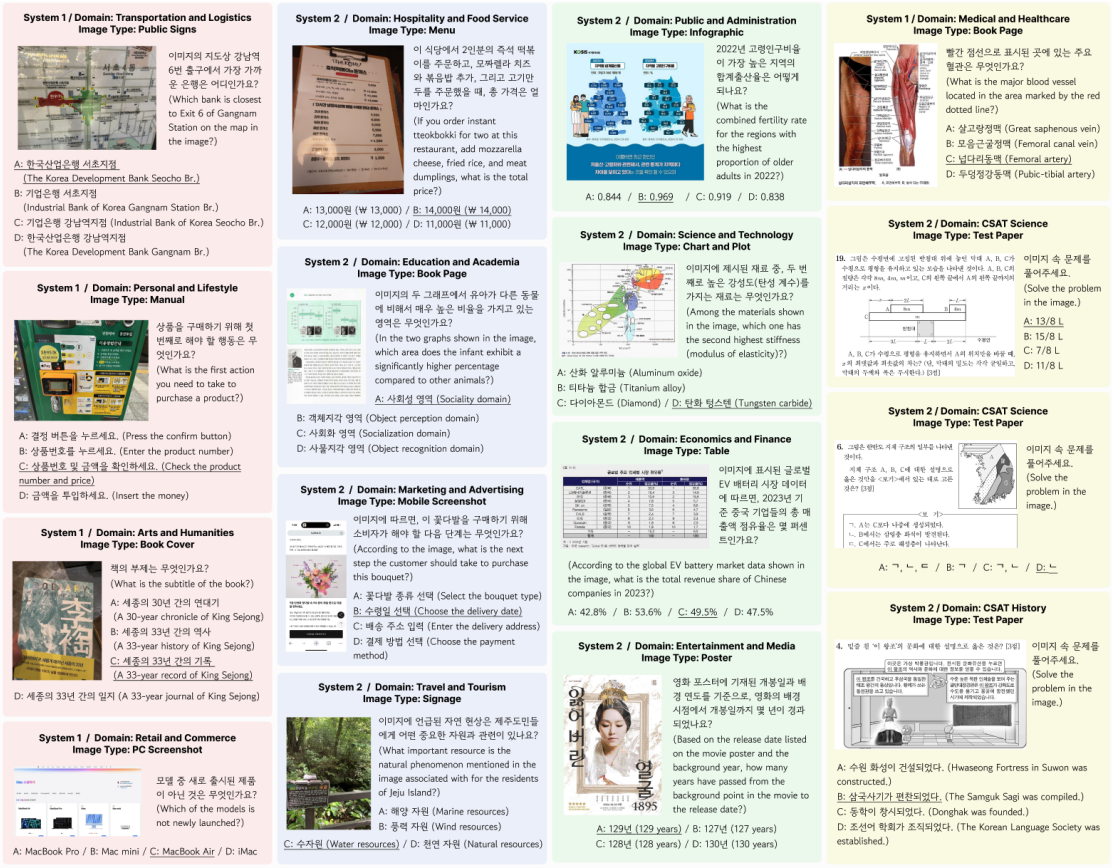

Figure 2: Examples from KRETA, showcasing diverse domains and image types categorized under System 1 and System 2. The model input consists of an image, a Korean question, and multiple-choice options.

2 Related Work

2.1 Vision-Language Models

Recent advancements in VLMs (Bai et al., 2025; Abdin et al., 2024; Chen et al., 2024b; Wu et al., 2024) have broadened their capabilities beyond traditional computer vision tasks, enabling contextual reasoning across visual domains and deeper language-vision integration. However, general-purpose VLMs often struggle with text-rich images, as they focus on holistic scene interpretation rather than precise text comprehension. To address this, text-centric VLMs such as LLaVAR (Zhang et al., 2024b), LLaVA-Read (Zhang et al., 2024a), and TextSquare (Tang et al., 2024a) enhance reading abilities by refining text recognition and reasoning. While these models improve performance on text-rich tasks, they are still English-only, highlighting the need for multilingual VLMs.

2.2 Text-Rich VQA Benchmarks

General VQA benchmarks (Lu et al., 2022; Yue et al., 2024a; Liu et al., 2023b) evaluate broad reasoning skills. However, benchmarks dedicated to text-rich VQA remain scarce, especially outside English. Early work such as TextVQA (Singh et al., 2019) and OCR-VQA (Mishra et al., 2019) targets printed English text (e.g., billboards, book covers). Moving beyond English, MTVQA (Tang et al., 2024b) provides multilingual annotations but is limited in scale, whereas MUST-VQA (Vivoli et al., 2022) expands data via automatic translation at the cost of language-specific nuance. xGQA (Pfeiffer et al., 2022) also relies on machine translation with only a single difficulty tier, and SEA-VQA (Urailertprasert et al., 2024) narrows its scope to Southeast-Asian heritage imagery. Meanwhile, KOFFVQA (Kim and Jung, 2025), though rule-based, remains small and still merges reading with reasoning. Most text-oriented VQA benchmarks favor high-resource languages or depend on translated English datasets, with limited support for Korean (Sun et al., 2024; Yue et al., 2024b).

3 KRETA Benchmark

As shown in Table 1, the KRETA benchmark is carefully designed to evaluate the ability of VLMs

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

Table 1: Comparison of Korean VQA Benchmarks. The Image Source column indicates native Korean images (Ko) or English-translated ones (En). Text-Centric Reasoning indicates whether the benchmark focuses on reasoning over text in images. Forms lists the answer type open-ended (Open), short-answer (Short), or multiple-choice (MC). Image Type categorizes images as General (non text-centric), Document (structured layouts), or Multi-text (diverse text-rich contexts).

to understand and reason about Korean text appearing in images. Importantly, rather than relying on translated English resources, all images and QA pairs in KRETA were originally generated in Korean, ensuring natural language usage and cultural relevance. To the best of our knowledge, with 2,577 samples, it stands among the largest Korean text-rich VQA datasets to date. As illustrated in Figure 2, KRETA features diverse visual contexts and requires advanced reasoning like domain-knowledge and multi-step cross-modal reasoning. The following subsections detail the dataset statistics and categorization, the data collection process, the semi-automated VQA generation pipeline, and the human annotation refinement process.

3.1 Data Statistics and Categorization

Our benchmark consists of 2,577 samples, each annotated with corresponding QA pairs. Each image is categorized into one or both reasoning levels: *System 1* (basic recognition and understanding) and *System 2* (advanced reasoning). In total, the dataset includes 1,426 System 1 QA pairs and 1,151 System 2 QA pairs. Beyond the in-depth analysis provided by the reasoning-based categorization, we conduct a multi-faceted analysis of VLM performance by categorizing images along two additional dimensions: *Domain* and *Image Type*. The images cover 26 distinct types across 15 domains.

System 1 vs. System 2 To assess challenges in visual text understanding and provide a comprehensive evaluation, we adopt a two-tiered cognitive framework (Kahneman, 2011; Yu et al., 2024) that distinguishes basic recognition (System 1, fast thinking) from advanced reasoning (System 2, slow thinking). System 1 relies on intuitive and automatic recognition, requiring direct text extraction and straightforward interpretation. In contrast, System 2 demands advanced reasoning, such as contextual understanding, multi-step decision-making, numerical reasoning, and integration of external knowledge when necessary.

Domain To ensure that our domain classification aligns with real-world industrial applications, we refer to the Korean Standard Industrial Classification (KSIC) (Korea Statistics, 2024) framework. We adapt this framework to suit our image data analysis, following a structured approach similar to MMMU (Yue et al., 2024a). We define 13 primary domains: Public & Administration (Gov.), Economics & Finance (Econ.), Marketing & Advertising (Mktg.), Retail & Commerce (Comm.), Education & Academia (Edu.), Medical & Healthcare (Med.), Science & Technology (Tech.), Arts & Humanities (Arts.), Transportation & Logistics (Transp.), Travel & Tourism (Tour.), Hospitality & Food Service (FnB.), Entertainment & Media (Ent.), and Personal & Lifestyle (Life.).

In addition, we incorporate CSAT (College Scholastic Ability Test) Science (Sci.) and History (Hist.) as separate domains. Unlike other domains generated by our semi-automated pipeline, CSAT questions were directly adapted from official examination materials. For each item, we crop the image region containing the question stem and its associated visual context, then supply the multiple-choice options to the model as text.

Image Type Images are categorized based on their inherent visual structures and the way they convey information. To systematically analyze VLM performance across different visual formats, we classify all images into 26 distinct types, ranging from highly structured (e.g., tables, receipts) to visually complex (e.g., posters, PC screenshots). These include charts and plots, infographics, posters, mobile/PC screenshots, manuals, receipts, street signs, menus, and more.

3.2 Data Collection

For this study, we compile images from copyright-free online repositories and our own field photography. To ensure balanced coverage of real-world scenarios, we identify domain imbalances and add samples in underrepresented categories. We source data from government publications, posters, free

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 (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**

Table 2: Evaluation results of closed and open-source VLMs on the KRETA, highlighting performance under the System 1 and System 2 framework. As marked in color, models struggle with System 2 reasoning tasks.

image databases, administrative documents, statistical reports from public agencies, and publicly available Korean mock exams including the CSAT.

3.3 Semi-Automated VQA Generation Pipeline

Step 1: Stepwise Image Decomposition In this step, we refine the dataset by filtering out low-quality images. Images with a shortest side of 384 pixels or less are discarded to ensure text readability. To further ensure meaningful textual content, we use PaddleOCR[1] to exclude images with fewer than 10 or more than 1,000 Korean characters.

Following the filtering process, multiple VLMs independently extract both textual and non-textual elements from each image. By default, we employ two foundation models, GPT-4o-mini (OpenAI, 2024a) and Gemini-2.0-flash (DeepMind, 2025), and merge their outputs to maximize extraction thoroughness while minimizing hallucinations. The structured decomposition process first analyzes non-textual visual attributes such as the overall scene, document layout, key objects, and background details. It then examines the structural and semantic relationships between text and visual components before finally extracting and structuring all textual content. This approach preserves contextual links between visual and textual elements, yielding higher-quality outputs than direct OCR alone.

[1] https://github.com/PaddlePaddle/PaddleOCR

Step 2: QA Candidates Generation Using the structured captions from Step 1, this step simultaneously generates question-answer candidates via LLMs. QA generation follows the System 1 and System 2 framework, with prompts specifically designed to assess different levels of visual text understanding and reasoning. For System 1, we use GPT-4o-mini (OpenAI, 2024a) and Gemini-2.0-flash (DeepMind, 2025) to generate two candidates each. For System 2, we employ o1-mini (OpenAI, 2024b) and Gemini-2.0-flash (DeepMind, 2025) to leverage their strong reasoning performance. The pipeline offers flexible control over the choice of model and the number of QA candidates generated.

Independently, the classification step assigns each image to its appropriate domain and image type as defined in Section 3.1, based on the structured captions from Step 1.

Step 3: QA Evaluation and Voting In this step, multiple VLMs (by default GPT-4o-mini (OpenAI, 2024a) and Gemini-2.0-flash (DeepMind, 2025)) evaluate the generated QA candidates to determine the highest-quality question-answer pair for each image. Drawing inspiration from prior LLM evaluation research (Zheng et al., 2023; Fu et al., 2024), the process employs a set of predefined criteria to systematically assess candidate quality.

For System 1 candidates, we use five metrics (Text Utilization, Clarity, Correctness, Naturalness, and Alignment) to ensure textual content accuracy and coherence. For System 2 candidates, two additional metrics (Complexity and Coherence) account for multi-step reasoning and logical inference. Each VLM assigns a score from 0 to 5 for each metric, and we use the aggregated scores to rank the candidates. A voting mechanism then selects the highest-ranked QA pair across all VLMs.

Step 4: Hard Negatives Generation After selecting the final QA pair, an LLM generates three hard negative options that resemble the correct answer while remaining distinct in meaning. These options follow the correct answer's structure and context, making the multiple-choice format more challenging.

Human Annotation Refinement The final QA pairs undergo a thorough human review based on the same evaluation criteria as Step 3. We adjust or remove questions that can be answered solely from text without image context (Text Utilization); verify that each QA pair aligns with the image's

Model	Size	Overall (2,577)	Gov. (245)	Econ. (104)	Mktg. (145)	Comm. (154)	Edu. (215)	Med. (90)	Tech. (92)	Arts. (83)	Transp. (167)	Tour. (108)	FnB. (264)	Ent. (168)	Life. (204)	*Sci.* (478)	*Hist.* (60)
Closed																	
GPT-4o (OpenAI, 2024a)	-	84.6	93.5	92.3	97.2	90.3	**96.7**	91.1	**96.7**	**100.0**	84.4	93.5	**93.6**	**97.0**	95.1	**44.1**	**93.3**
GPT-4o-mini (OpenAI, 2024a)	-	73.3	82.4	82.7	85.5	84.4	87.4	83.3	80.4	89.2	80.2	84.3	81.4	86.3	87.3	30.3	45.0
Gemini-2.0-flash (DeepMind, 2025)	-	**85.4**	**95.1**	**95.2**	**99.3**	**96.1**	**96.7**	**92.2**	93.5	98.8	**90.4**	**98.1**	93.2	95.2	**96.6**	**44.1**	78.3
Claude-3.5-Sonnet (Anthropic, 2024)	-	80.5	93.5	91.3	92.4	87.0	93.0	91.1	87.0	91.6	84.4	94.4	89.8	92.3	92.2	37.4	70.0
Open-source																	
LLaVA-OneVision (Li et al., 2024)	0.5B	42.3	51.8	48.1	47.6	44.8	39.5	50.0	44.6	40.9	49.7	51.9	41.7	44.6	46.1	28.0	31.7
Deepseek-VL2-tiny (Wu et al., 2024)	1B	48.8	57.1	55.8	63.4	58.4	51.2	57.8	57.6	45.8	54.5	58.3	43.9	47.0	54.4	30.5	31.7
Deepseek-VL2-small (Wu et al., 2024)	2.8B	53.3	61.6	63.5	66.9	63.0	57.2	64.4	68.5	50.6	59.9	63.0	48.9	56.0	57.4	30.8	36.7
Qwen2.5-VL (Wang et al., 2024)	3B	**71.8**	81.6	**76.9**	85.5	77.9	**87.4**	**80.0**	**79.3**	**85.5**	**75.4**	**84.3**	**76.9**	**87.5**	83.3	**33.9**	36.7
Ovis1.6-Llama3.2 (Lu et al., 2024)	3B	52.2	64.5	69.2	60.7	57.1	55.8	54.4	62.0	51.8	60.5	61.1	56.8	52.4	49.5	30.5	31.7
InternVL2.5 (Chen et al., 2024b)	4B	70.7	**82.0**	**76.9**	**87.6**	**83.1**	83.7	78.9	**79.3**	79.5	75.4	77.8	69.3	81.0	**86.3**	**33.9**	**46.7**
Phi-3.5-Vision (Abdin et al., 2024)	4.2B	42.6	53.5	55.8	40.0	49.4	43.3	40.0	53.3	50.6	44.3	46.3	42.8	43.5	44.6	27.6	36.7
LLaVA-OneVision (Li et al., 2024)	7B	54.0	64.1	63.5	63.4	63.6	58.6	55.6	64.1	45.8	68.3	65.7	55.3	55.4	55.9	30.8	33.3
Qwen2.5-VL (Wang et al., 2024)	7B	68.5	80.0	77.9	**85.5**	81.2	**87.4**	76.7	75.0	**89.2**	77.8	82.4	**77.7**	**86.3**	**85.8**	15.1	36.7
InternVL2.5 (Chen et al., 2024b)	8B	70.8	**81.6**	76.9	**85.5**	81.8	83.7	81.1	77.2	78.3	76.0	**83.3**	74.2	78.6	**85.8**	**34.1**	**38.3**
MiniCPM-V-2.6 (Yao et al., 2024)	8B	41.0	50.2	54.8	50.3	53.2	44.7	41.1	52.2	33.7	43.7	48.1	43.6	45.8	46.1	18.2	25.0
MiniCPM-o-2.6 (Yao et al., 2024)	8B	64.3	75.9	83.7	79.3	75.9	76.7	65.6	75.0	73.5	69.5	79.6	67.8	77.4	74.0	25.5	25.0
Ovis1.6-Gemma2 (Lu et al., 2024)	9B	58.4	64.1	69.2	71.0	72.7	60.9	71.1	67.4	53.0	68.9	75.9	65.2	58.9	63.2	30.5	28.3
VARCO-VISION (Ju et al., 2024)	14B	**72.3**	**81.6**	**87.5**	83.4	**83.1**	84.2	**86.7**	**84.8**	79.5	**82.6**	**83.3**	76.1	81.5	85.3	33.7	31.7

Table 3: Evaluation results for closed and open-source VLMs on KRETA across 15 domains.

original intent (Alignment); confirm that System 2 questions require at least one inferential step to avoid overly simple QA (Complexity); and review language, grammar, and factual content (Naturalness, Correctness, and Clarity).

4 Empirical Analysis

We leverage VLMEvalKit (Duan et al., 2024), an open-source evaluation toolkit designed to facilitate the assessment of VLMs, including both proprietary APIs and open-source models. We adopt the multiple-choice system prompt from MMMU-Pro (Yue et al., 2024c). The prompt instructs the model as follows: *Please select the correct answer from the options above. The last line of your response should follow the format: 'Answer: LETTER' (without quotes), where LETTER corresponds to one of the provided options.*

4.1 Performance across System 1 vs. System 2

Table 2 presents the performance breakdown between System 1 and System 2. Across both open-source and closed models, System 1 accuracy is significantly higher, indicating that most models handle text recognition and simple contextual understanding well. Notably, Gemini-2.0-flash (DeepMind, 2025) achieves 98.0% on System 1, reflecting near-perfect perception.

However, System 2 results reveal substantial performance drops, particularly in open-source models. Qwen2.5-VL-7B (Wang et al., 2024) falls from 94.5% in System 1 to 36.1% in System 2, and Deepseek-VL2-small (Wu et al., 2024) drops from 67.3% to 36.1%. GPT-4o (OpenAI, 2024a) retains a relatively stronger System 2 performance at 70.5%, yet this value remains suboptimal. Both open-source and closed models struggle in System 2, as effective reasoning requires sequential integration of multiple visual and textual cues, a capability that is still underdeveloped. These challenges are compounded by the low-resource nature of Korean pretraining, and by gaps in domain-specific and cultural knowledge, since models have limited exposure to Korean-contextualized data during training.

4.2 Performance across Domain

Table 3 compares closed and open-source model performance across 15 domains. Among closed models, Gemini-2.0-flash (DeepMind, 2025) achieves the highest overall score (85.4%), followed by GPT-4o (84.6%). Notably, GPT-4o excels in the CSAT History domain with 93.3%, suggesting strong historical and cultural reasoning. Gemini-2.0-flash's consistently high performance across domains further reflects its robust text recognition and contextual comprehension on real-world images.

Open-source models exhibit a broad range of performance on KRETA, with overall scores varying from 42.3% (LLaVA-OneVision (Li et al., 2024), 0.5B) to 72.3% (VARCO-VISION (Ju et al., 2024), 14B). The strongest performers are Qwen2.5-VL (Wang et al., 2024), InternVL2.5 (Chen et al., 2024b), and VARCO-VISION, each scoring in the low 70% range. Notably, Qwen2.5-VL (7B) maintains high accuracy in practical domains such as Marketing (85.5%) yet plunges to 15.1% in CSAT Science. Applying Chain-of-Thought prompting (Section 5) substantially boosts Qwen2.5-VL's Sys-

Image Type	Closed		Open		Sys1 - Sys2		Closed - Open	
	Sys1	Sys2	Sys1	Sys2	Closed	Open	Sys1	Sys2
Document								
Chart and Plot	94.9	86.7	79.3	48.2	8.2	31.1	15.6	38.5
Table	91.0	75.0	70.9	42.3	**16.0**	28.6	20.1	32.7
Infographic	95.4	81.3	**80.0**	44.1	14.1	**35.9**	15.4	37.2
Slides	96.4	**95.0**	73.0	**61.3**	1.4	11.7	23.4	33.7
Book Cover	95.4	91.0	69.0	52.0	4.4	17.0	**26.4**	**39.0**
Product Detail	94.3	87.5	78.6	51.5	6.8	27.1	15.7	36.0
Poster	94.6	87.3	73.8	54.0	7.3	19.8	20.8	33.3
Mobile Screen	**97.2**	90.7	76.9	54.9	6.5	22.0	20.3	35.8
PC Screen	94.8	83.6	74.8	50.1	11.2	24.7	20.0	33.5
Scene Text								
Street Signs	87.0	**93.1**	75.9	**59.3**	-6.1	16.6	11.1	33.8
Public Signs	88.6	69.4	71.2	42.0	19.2	29.2	17.4	27.4
Store Sign	91.4	85.3	70.6	42.0	6.1	28.6	20.8	43.3
Banner	94.6	91.1	78.2	46.2	3.5	**32.0**	16.4	**44.9**
Signage	**94.7**	85.9	**78.5**	54.3	8.8	24.2	16.2	31.6
Menu	91.9	79.9	69.5	40.3	12.0	29.2	**22.4**	39.6
Manual	91.2	71.1	73.2	42.1	**20.1**	31.1	18.0	29.0

Table 4: Performance comparison across image types for closed and open-source models, showing differences across System 1, System 2, and model categories. Only image types with at least 50 VQA pairs are presented.

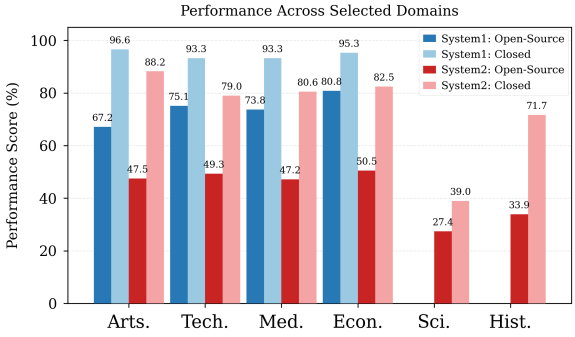

Figure 3: Comparison of open-source and closed models across different domains on KRETA. Bars show the average scores of closed and open-source models separately for System 1 and System 2 in each domain.

tem 2 performance overall, underscoring its original deficiency in multi-step reasoning and external knowledge integration. Taken together, Table 3 reveals significant variability among open-source models in both overall and domain-specific metrics, highlighting the need to carefully consider model size, architecture, and domain alignment when selecting a model for a given application.

Figure 3 illustrates the System 1 and System 2 performance gap between closed and open-source models across different domains on KRETA. The disparity is particularly pronounced in System 2 tasks, where closed models outperform open-source counterparts by up to 40.7 percentage points in *Arts & Humanities*, reflecting stronger cultural understanding. Meanwhile, the *Science & Technology* domain shows a relatively smaller System 2 gap of 29.7 percentage points, suggesting more consistent handling of technical content. In the CSAT domains, gaps of 11.6 points in Science and 37.8 points in History further underscore the role of background knowledge. These findings suggest that open-source models need targeted domain-specific training, particularly in culturally and historically rich areas, to close the reasoning gap with closed models.

4.3 Performance across Image Type

Table 4 presents the performance of closed and open-source models across different image types, highlighting key trends in System 1 and System 2 tasks. Performance varies significantly by image type, reflecting distinct model capabilities. Document-based images such as tables and infographics achieve high accuracy for closed-source models in System 1 (91.0%, 95.4%) and retain relatively strong performance in System 2 (75.0%, 81.3%). In contrast, open-source models fall to 42.3% on Tables and 44.1% on Infographics in System 2. Notably, Book Covers exhibit the largest closed-open gap: 26.4 points in System 1 and 39.0 points in System 2, likely due to their complex typography and mixed visual elements.

Scene-text images present different challenges. Street Signs show a rare pattern for closed models, with System 2 accuracy (93.1%) exceeding System 1 (87.0%), possibly because motion blur or low resolution impairs simple text extraction while System 2 can leverage broader context. In contrast, open-source models perform particularly poorly on banners and store signs, where the System 2 gap reaches 44.9 points and 43.3 points, respectively, indicating difficulties with diverse fonts, occlusions, and unconventional layouts common in real-world signage. These findings highlight the varying complexity of image types and underscore the need for targeted improvements in both structured-text processing and robust scene-text understanding.

4.4 Performance across Closed vs. Open-source

Overall, closed-source models outperform open-source counterparts by an average of 24.2 percentage points in overall score, with the System 2 reasoning gap reaching as high as 44.4 percentage points, revealing a pronounced reasoning bottleneck in open models (Table 2). Domain analysis shows a relatively modest closed-open gap of 23.7 percentage points in the *Science & Technology* do-

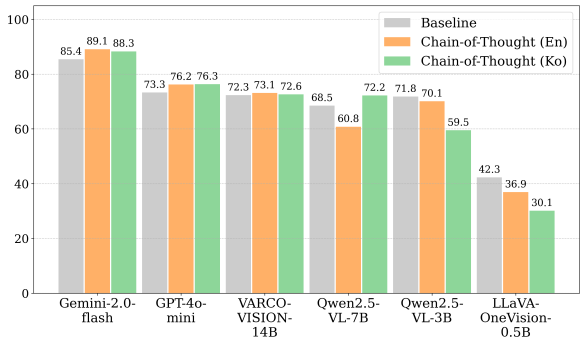

Figure 4: Comparison of two closed and four open-source models of varying sizes on KRETA. The figure shows performance differences across three prompts: Baseline, Chain-of-Thought in English and Korean.

main, but this difference widens to 40.7 percentage points in CSAT History, highlighting the closed models' superior ability to integrate background knowledge and cultural context (Table 3). Similarly, across image types, from structured documents such as tables and infographics to cluttered, unstructured layouts such as banners and signage, the transition from closed to open models yields comparable performance declines, underscoring open models' limited versatility in handling diverse visual-textual presentations (Table 4).

4.5 Performance across Model Size

Table 3 demonstrates a clear positive correlation between model capacity and overall performance: Deepseek-VL2 improves from 48.8 at 1 B parameters (tiny) to 53.3 at 2.8 B (small), and LLaVA-OneVision rises from 42.3 at 0.5 B to 54.0 at 7 B. These results confirm that, for a given architecture, increasing model size generally yields gains in both aggregate accuracy and domain-specific metrics. An exception to this trend is observed with Qwen2.5-VL, where the 3 B variant (71.8) outperforms the 7 B variant (68.5). This anomaly suggests that the addition of further multilingual data during scaling may have diluted the model's Korean-centric knowledge and reasoning abilities. Consequently, when enlarging multilingual VLMs, it is essential to preserve the proportion of low-resource language data and to apply domain-adaptive fine-tuning to sustain performance on language-specific and culturally nuanced tasks.

5 Discussion

Chain-of-Thought (CoT) We evaluate the impact of Chain of Thought (CoT) prompting on model performance, following the approach demonstrated in MMMU-Pro (Yue et al., 2024c). Figure 4 reveals a pronounced gap between closed and open-source models in both baseline scores and CoT improvements.

Closed models benefit consistently. For instance, Gemini 2.0-flash improves by 3.7 points with English CoT and by 2.9 points with Korean CoT, indicating robust instruction following and structured reasoning. Mid-size open-source models exhibit language-dependent effects. Qwen2.5-VL-7B declines by 7.7 points with English CoT but improves by 3.7 points with Korean CoT, suggesting sensitivity to prompt language and potential for language-specific optimization. Lightweight open-source models suffer performance degradation under CoT prompting. LLaVA-OneVision-0.5B drops by 5.4 points under English CoT and by 12.2 points under Korean CoT. These declines suggest that excessive reasoning instructions overwhelm models with limited capacity.

Overall, these results demonstrate that CoT prompting enhances performance only when a model possesses adequate reasoning capacity and instruction-following ability, and may become detrimental otherwise.

6 Conclusion

In this paper, we present KRETA, a comprehensive benchmark for evaluating VLMs on Korean text-rich images. KRETA adopts a dual-level reasoning framework for both basic recognition and advanced inference, and employs an industry-aligned domain and image-type taxonomy spanning 15 domains and 26 image formats. By relying exclusively on native Korean imagery and questions, our benchmark extends beyond previous Korean or multilingual VQA sets that have been confined to document-only tasks or machine-translated content. Our training-free, semi-automated pipeline combines structured image decomposition with cross-validation by two foundation models and a novel multi-metric evaluation protocol. Our experimental results underscore the need for domain-adaptive fine-tuning, the careful preservation of low-resource language data balance during scaling, and the integration of stronger reasoning mechanisms. We hope that our VQA generation pipeline will be readily transferable to other low-resource languages, laying the groundwork for culturally and linguistically tailored VLMs.

Limitations

While we provide a comprehensive evaluation of Korean text-rich VQA, several limitations suggest directions for future work. First, KRETA is confined to single-image, multiple-choice question answering. Extending the benchmark to include multi-image or video-based scenarios, and to incorporate high-level comprehension tasks (e.g. section-to-section verification, information synthesis, document summarization, open-ended generation), would yield a more complete assessment of vision-language capabilities.

Second, the System 2 category conflates sequential deduction, integration of external information and cross-referential contextual analysis into a single classification. Developing a more fine-grained taxonomy to distinguish these reasoning functions would expose specific model weaknesses and support targeted improvements. In this work, we prioritized the creation of a scalable, high-quality unified benchmark spanning 15 domains and 26 image formats, establishing a solid foundation for Korean text-rich VQA while leaving finer reasoning taxonomies and additional task formats to future extensions.

Lastly, Chain-of-Thought (CoT) prompting has been shown to improve performance on the System 2 benchmark, particularly for closed-source models that consistently gain from both English and Korean CoT variants, but additional strategies and prompt formulations remain unexplored. Investigating alternative CoT techniques, hybrid reasoning frameworks, and other optimization methods for both closed-source and open-source models represents an open challenge for future research. We hope that KRETA serves as a stepping stone for future advancements in this area, guiding the development of more effective reasoning strategies and robust VLMs.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Anthropic. 2024. The Claude 3 model family: Opus, sonnet, haiku. https://www.anthropic.com/news/claude-3-5-sonnet. Claude-3.5-sonnet-2024-10-22 version [Multimodal Large Language Model].

Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration. *arXiv preprint arXiv:2406.16469*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. 2024a. MJ-Bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

DeepMind. 2025. Gemini 2.0 Flash main page. https://deepmind.google/technologies/gemini/flash. Gemini-2.0-flash-exp version [Multimodal Large Language Model].

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pages 11198–11201.

Weiping Fu, Bifan Wei, Jianxiang Hu, Zhongmin Cai, and Jun Liu. 2024. QGEval: A benchmark for question generation evaluation. *arXiv preprint arXiv:2406.05707*.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.

Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. 2022. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv preprint arXiv:2209.08199*.

Jeongho Ju, Daeyoung Kim, SunYoung Park, and Youngjune Kim. 2024. VARCO-VISION: Expanding frontiers in korean vision-language models. *arXiv preprint arXiv:2411.19103*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, NY, USA.

Yoonshik Kim and Jaeyoon Jung. 2025. Koffvqa: An objectively evaluated free-form vqa benchmark for large vision-language models in the korean language. *arXiv preprint arXiv:2503.23730*.

Korea Statistics. 2024. KSIC: The korean standard industrial classification, january, 2024. https://classification.codes/classifications/industry/ksic.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal reasoning via thought chains for science question answering. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. OVIS: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual question answering by reading text in images. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.

OpenAI. 2024a. GPT-4o main page. https://openai.com/index/hello-gpt-4o. Gpt-4o-2024-11-20 version [Multimodal Large Language Model].

OpenAI. 2024b. o1-mini main page. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/. O1-mini-2024-09-12 version [Large Language Model].

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. 2024. Parrot: Multilingual visual instruction tuning. *arXiv preprint arXiv:2406.02539*.

Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, et al. 2024a. TextSquare: Scaling up text-centric visual instruction tuning. *arXiv preprint arXiv:2404.12803*.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024b. MTVQA: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185.

Emanuele Vivoli, Ali Furkan Biten, Andres Mafla, Dimosthenis Karatzas, and Lluis Gomez. 2022. MUST-VQA: Multilingual scene-text vqa. *arXiv preprint arXiv:2209.06730*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. 2024. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024a. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024b. Pangea: A fully open multilingual multimodal LLM for 39 languages. *arXiv preprint arXiv:2410.16153*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2024c. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Ruiyi Zhang, Yufan Zhou, Jian Chen, Jiuxiang Gu, Changyou Chen, and Tong Sun. 2024a. LLaVA-Read: Enhancing reading ability of multimodal language models. *arXiv preprint arXiv:2407.19185*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2024b. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:46595–46623.