Cross-Lingual Text-Rich Visual Comprehension: An Information Theory Perspective

Xinmiao Yu[1], Xiaocheng Feng[1]*, Yun Li[1], Minghui Liao[2], Ya-Qi Yu[2], Xiachong Feng[3], Weihong Zhong[1], Ruihan Chen[1], Mengkang Hu[3], Jihao Wu[2], Duyu Tang[2], Dandan Tu[2], Bing Qin[1]*

[1]Harbin Institute of Technology
[2]Huawei Inc.
[3]The University of Hong Kong
{xmyu, xcfeng, yunli, whzhong, rhchen}@ir.hit.edu.cn, fengxc@hku.hk, mkhu@connect.hku.hk,
{liaominghui1, yuyaqi5, wujihao, tangduyu, tudandan}@ir.hit.edu.cn

Abstract

Recent Large Vision-Language Models (LVLMs) have shown promising reasoning capabilities on text-rich images from charts, tables, and documents. However, the abundant text within such images may increase the model's sensitivity to language. This raises the need to evaluate LVLM performance on cross-lingual text-rich visual inputs, where the language in the image differs from the language of the instructions. To address this, we introduce **XT-VQA** (**C**ross-Lingual **T**ext-Rich **V**isual **Q**uestion **A**nswering), a benchmark designed to assess how LVLMs handle language inconsistency between image text and questions. XT-VQA integrates five existing text-rich VQA datasets and a newly collected dataset, XPaperQA, covering diverse scenarios that require faithful recognition and comprehension of visual information despite language inconsistency. Our evaluation of prominent LVLMs on XT-VQA reveals a significant drop in performance for cross-lingual scenarios, even for models with multilingual capabilities. A mutual information analysis suggests that this performance gap stems from cross-lingual questions failing to adequately activate relevant visual information. To mitigate this issue, we propose **MVCL-MI** (**M**aximization of **V**ision-Language **C**ross-**L**ingual **M**utual **I**nformation), where a visual-text cross-lingual alignment is built by maximizing mutual information between the model's outputs and visual information. This is achieved by distilling knowledge from monolingual to cross-lingual settings through KL divergence minimization, where monolingual output logits serve as a teacher. Experimental results on the XT-VQA demonstrate that MVCL-MI effectively reduces the visual-text cross-lingual performance disparity while preserving the inherent capabilities of LVLMs, shedding new light on the potential practice for improving LVLMs.

Introduction

Large Vision-Language Models (LVLMs) have achieved significant advancements in domains such as mathematical reasoning (Lu et al. 2023), multimodal search (Yang et al. 2024) and embodied intelligence (Mu et al. 2024). Notably, their robust multimodal capabilities demonstrate superiority in handling text-rich scenarios, leading to various applications, including document processing (Luo et al. 2024), free-form web auto-manipulation (Niu et al. 2024), and scene text understanding (Yu et al. 2024; Ye et al. 2023). These works meticulously explore the ability of LVLMs to recognize, process, and analyze multimodal information based on instructions, yielding notable progress.

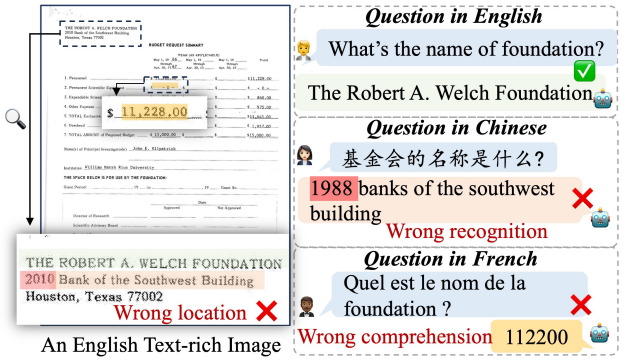

Figure 1: An example of the LVLM answering unfaithfully when questions were posed in languages different from those in the image. The LVLM made unfaithful recognition and comprehension of Chinese and French while answering correctly with English questions. Reveals the challenge of cross-lingual visual comprehension.

However, current research on text-rich visual comprehension primarily focuses on monolingual settings, largely neglecting the performance of LVLMs in cross-lingual scenarios where the instruction language differs from the textual language in the visual content, (see Fig 1.) This gap limits many real-world applications. For example, in a *foreign airport with signs in unfamiliar language*, the ability to query an LVLM *in your native language* for assistance would be invaluable. As globalization accelerates, cross-lingual scenarios will become increasingly common across domains such as healthcare (Wan et al. 2024), law (Guha et al. 2024), and science (Lu et al. 2022). Investigating the cross-lingual instruction-following capabilities of LVLMs (Hinck et al.

*Corresponding author

2024) is therefore essential. To address this, our work systematically explores the task of cross-lingual text-rich visual comprehension by tackling three key scientific questions.

First, we address the question: *"Does the cross-lingual scenario impact the text-rich visual comprehension capabilities of LVLMs?"* To answer this, we construct the XT-VQA (**Cross**-Lingual **Text**-Rich **V**isual **Q**uestion **A**nswering) benchmark to overcome the challenge of data scarcity. XT-VQA integrates multiple existing VQA datasets (Mathew, Karatzas, and Jawahar 2021; Masry et al. 2022; Singh et al. 2019; Mishra et al. 2019) and introduces the newly curated XPaperQA dataset, which focuses on bilingual academic papers. Designed to study cross-lingual text-rich visual comprehension, XT-VQA covers diverse visual information types, including charts, scene text, and documents. XPaperQA, a key component of XT-VQA, contains 4,436 question-answer pairs generated using the advanced Gemini-Pro model, with rigorous filtering and quality review processes ensuring high data quality. Notably, XPaperQA addresses the scarcity of non-English images in existing datasets. Experimental results on XT-VQA reveal that while LVLMs demonstrate multilingual capabilities, they face significant difficulties in cross-lingual text-rich visual comprehension, with performance dropping by 32.6%.

Next, we address the second question: *"What causes the performance decline of LVLMs in cross-lingual text-rich visual comprehension scenarios?"* Inspired by prior work leveraging information theory to analyze performance gaps (Farquhar et al. 2024), we examine the performance drop on XT-VQA from an information-theory perspective. Since answers in XT-VQA are typically embedded in textual form within images, effective comprehension of visual information is crucial for LVLMs to perform well. To quantify the role of visual information across languages, we analyze the mutual information between the model output and the input image. Our analysis reveals a strong correlation between accuracy and mutual information, suggesting that increasing mutual information between the visual and language components could mitigate the cross-lingual performance gap.

Finally, we address the third question: *"How can we mitigate this gap while retaining monolingual capability?"* To this end, we propose **MVCL-MI** (**M**aximize **V**ision-Language **C**ross-**L**ingual **M**utual **I**nformation), a method designed to enhance the activation of visual information in LVLMs. MVCL-MI improves cross-lingual performance on XT-VQA while preserving monolingual capabilities by leveraging cross-lingual distillation to maximize mutual information between visual and language modalities across different languages. We evaluate MVCL-MI on the XT-VQA benchmark, comparing it with existing LVLMs. Experimental results show that MVCL-MI effectively reduces the performance gap in cross-lingual settings while maintaining strong monolingual performance. Ablation studies further confirm that the improvements in accuracy and fidelity stem from enhanced mutual information across modalities and languages.

Related Works

Text-Rich Multimodal Understanding Text-rich multimodal understanding requires VLMs' abilities to recognize, understand, and reason over the text content contained in images (Mathew, Karatzas, and Jawahar 2021; Masry et al. 2022; Mishra et al. 2019; Singh et al. 2019). Many works try to improve text-rich visual comprehension. CLIPPO (Tschannen, Mustafa, and Houlsby 2023) further improves the CLIP (Radford et al. 2021) by training with image and rendered text pair alignment. Pix2Struct (Lee et al. 2023) trains a powerful end2end model to convert text-rich screenshots into structural HTML code. As the development of instruction fine-tuning in LLM (Brown et al. 2020; Touvron et al. 2023), LVLM uses the projector to align visual tokens to text tokens and then does visual instruction tuning based on the LLM backbone (Liu et al. 2023; Bai et al. 2023; Dai et al. 2023). Works (Li et al. 2024; Yu et al. 2024) further enriches the instruction tuning dataset with OCR data results in OCR-related task performance rise remarkably.

Cross-lingual in Multimodal Cross-lingual research is a key area in natural language processing, covering tasks such as cross-lingual information retrieval, question answering, and summarization (Thakur et al. 2024; Wang et al. 2023; Chen et al. 2023; Huang et al. 2024). While prior studies have assessed the multilingual capabilities of LVLMs (Schneider and Sitaram 2024; Wang et al. 2024), including training models in specific languages such as Arabic (Andersland 2024) and English-Korean-Chinese trilingual models (Shin et al. 2024), their ability to handle cross-lingual tasks in visual contexts remains underexplored. While MTVQA (Tang et al. 2024) investigates same-language visual-text alignment in multilingual settings, our work uniquely focuses on cross-lingual inconsistencies in visual comprehension, specifically targeting real-world applications such as interpreting foreign signs.

Information theory in Multimodal Information theory interrelates with deep learning tightly. (Tishby and Zaslavsky 2015) employs information bottleneck as the theoretical framework for analyzing deep learning. Decoding approaches that leverage mutual information scores have demonstrated their usefulness across various scenarios (Li and Jurafsky 2016). For instance, they have proven beneficial in zero-shot settings (**?**) or when aiming to promote diversity and relevance in neural dialogue models (**??**) Mutual information has been used in alleviating hallucinations in language models (Xiao and Wang 2021). (Nandwani et al. 2023)Use conditional pointwise mutual information as score to quantify the faithfulness of models' response.

XT-VQA Benchmark

Problem Formulation

Formally, a cross-lingual text-rich question-answer pair can be represented as a text-rich image I containing text in a source language L^{src}, a question Q in target language L^{tgt}, where $L^{tgt} \neq L^{src}$. The goal is to accurately predict the answer A to the question Q, by effectively leveraging the visual and textual information present in image I, despite the language mismatch between L^{src} and L^{tgt}.

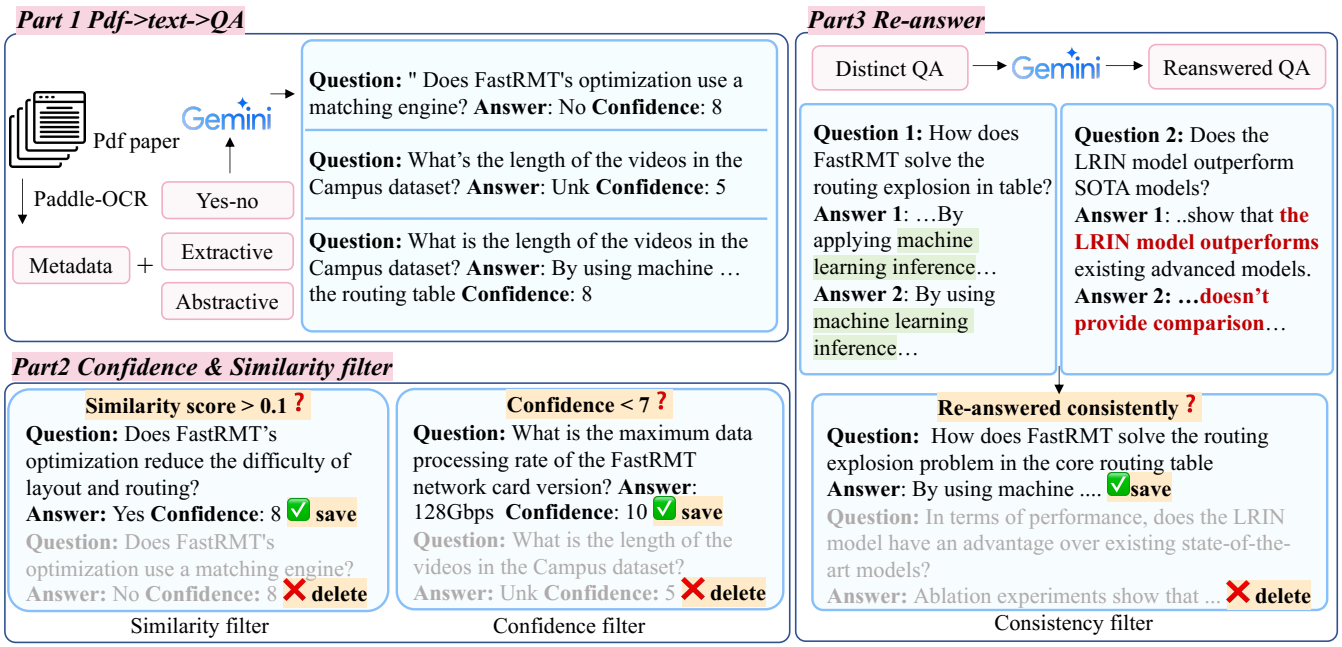

Figure 2: The XPaperQA dataset construction pipeline consists of three parts: (1) Converting PDF papers into metadata using PaddleOCR and generating three QA types via Gemini. (2) Filtering QA pairs with similarity scores > 0.1 or confidence scores < 7 to retain distinct pairs. (3) Re-answering the distinct QA pairs through Gemini and discarding inconsistent responses.

Dataset Construction

Our dataset construction has two parts. First, to evaluate the performance gap of LVLMs across scenarios—charts, documents, and scene text, we expand four established benchmarks into three languages. Second, since existing datasets primarily contain images with English text, it is unclear whether the gap arises from cross-language interference or the abundance of English multimodal training data. To address this, we developed XPaperVQA, a dataset with text-rich images in Chinese and English. Below, we detail the construction process.

Multilingual Text-rich VQA extension We use Google Translate[1] to extend existing text-rich Visual Question Answering datasets into multiple languages: English, Chinese, and French. To improve robustness and reduce translation biases, we apply back-translation and calculate BERT sentence similarity between the original and back-translated questions. Questions with similarities below a predefined threshold are manually corrected.

$$\theta = \sum_{(i,j)\in WordPairs} \max(cosSim(BERT_{emb}(tokX_i), BERT_{emb}(tokY_j))) \quad (1)$$

Here, $cosSim$ denotes the cosine similarity between the BERT embeddings of token i from the original sentence ($tokX_i$) and token j from the translated sentence ($tokY_j$).

[1]https://translate.google.com/?sl=en&tl=zh-CN&op=translate

For English papers, we reconstruct the QASPER dataset (Dasigi et al. 2021), which contains 5,049 questions across 1,585 Natural Language Processing papers, categorized into three types: extractive, abstractive, and yes-no. To adapt this text-only dataset for cross-lingual text-rich QA, we use PyMuPDF[2] to automatically extract structural metadata from the PDF files. For each question, we locate the page containing evidence to answer it. If such a page exists, we save it as a document image along with the QA pair; otherwise, we discard the pair. This filtering process results in 1,536 VQA pairs. For Chinese papers, we develop an automatic QA generation pipeline using papers collected from an authoritative Chinese computer science journal[3].

As shown in Figure 2, we split each paper's PDF into pages and use PaddleOCR[4] to extract text. The extracted text, combined with a QA generation prompt, is input into Gemini. Following (Dasigi et al. 2021), we design three academic QA types: yes-no, extractive, and abstractive. To ensure faithfulness and diversity, we apply strict filters. For accuracy, the model self-rates answers on a confidence scale of $1-10$, discarding pairs with scores below 7. For diversity, we remove pairs with Jaccard similarity exceeding 0.1:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Finally, to further improve the robustness, we ask Gemini to re-answer the generated question, we provide Gemini

[2]https://pymupdf.readthedocs.io/en/latest/
[3]http://jcip.cipsc.org.cn/CN/home
[4]https://github.com/PaddlePaddle/PaddleOCR/tree/main

Model	OCRVQA			Text-VQA			ChartVQA			DocVQA		
	en	zh	fr	en	zh	fr	en	zh	fr	en	zh	fr
Open-sourced												
LLaVA-v1.5-13b	**62.7***	29.1* -33.6	31.5* -31.2	**61.2***	5.9* -55.3	11.8* -49.4	**11.1**	5.8 -5.3	7.3 -3.8	**2.8**	1.3 -1.5	1.2 -1.6
LLaVA-v1.6-34b	**64.8**	48.1 -16.7	39.8 -25.0	**64.9**	54.5 -10.4	26.7 -38.2	**52.4**	33.7 -18.7	31.8 -20.6	**78.2**	61.4 -16.8	64.0 -14.2
InstructBLIP	**24.4**	10.9 -13.5	17.8 -6.6	**50.3***	34.5* -15.8	37.1* -13.2	**29.7**	21.6 -8.1	19.8 -9.9	**5.2**	3.1 -2.1	4.0 -1.2
mPlug-Owl2	**70.7**	65.0 -5.7	65.2 -5.5	**54.3**	45.4 -8.9	46.7 -7.6	**44.3**	23.2 -21.1	19.3 -25.0	**28.7**	21.0 -7.7	20.7 -8.0
Qwen-VL-Chat	**65.6**	36.1 -29.5	32.3 -33.3	**61.6**	35.1 -26.5	28.3 -33.3	**57.3**	45.9 -11.4	38.8 -18.5	**59.1**	26.9 -32.2	30.3 -28.8
Monkey	**70.4**	46.3 -24.1	48.9 -21.5	**61.6**	33.8 -27.8	35.8 -25.8	**64.6**	55.8 -8.8	52.9 -11.7	**65.9**	51.7 -14.2	49.8 -16.1
Cog-VLM	**70.5**	63.8 -6.7	61.9 -8.6	**78.9**	66.0 -12.9	66.1 -12.8	**57.6**	47.5 -10.1	48.9 -8.7	**65.4**	42.9 -22.5	45.5 -19.9
MiniCPM-V	**69.5**	46.9 -22.6	55.9 -13.6	**76.6**	63.5 -13.1	52.7 -23.9	**73.0**	62.9 -10.1	63.2 -9.8	**84.9**	71.6 -13.3	74.6 -10.3
GPT-4o	**52.3**	46.8 -5.5	46.8 -5.5	**72.6**	66.9 -5.7	65.4 -7.2	**69.1**	64.6 -4.5	64.3 -4.8	**74.7**	69.3 -5.4	68.2 -6.5
Gemini-1.5-flash	**55.9**	49.5 -6.4	49.5 -6.4	**72.7**	64.0 -8.7	70.7 -2.0	**69.1**	59.3 -9.8	63.1 -6.0	**76.1**	69.8 -6.3	70.4 -5.7

Table 1: LVLMs performance on XT-VQA. Accuracy of question in source language L^{src} are **bold**. 29.1 –33.6 indicates accuracy decrease compared to queries in source language L^{src}. The $*$ notes that auxiliary OCR tokens are used. Underline are used to mark the highest accuracy among LVLMs. For closed-sourced models, testing is conducted on a subset only.

with the edit distance as a reference to detect the answer consistency between the successive answers. At last, we obtain 3,870 QA pairs with 1,039 paper images.

$$d(a_1, a_2) = \min\{insert, delete, replace\} \quad (3)$$

	abstractive	extractive	yes-no	images
origin	8,289	12,241	14,359	1,199
+confidence filter	8,112	12,040	14,261	1,199
+similarity filter	1,374	1,530	1,687	1,074
+consistency filter	1,369	1,501	1,586	1,072
final	1,369	1,501	1,000	1,039

Table 2: Data statistics of XPaperQA

To ensure the quality and reliability of the XPaperQA dataset, we conducted a rigorous manual evaluation. We randomly sampled 100 questions from both the English and Chinese subsets and verified the correctness of their corresponding answers. The evaluation showed an accuracy of 87%, demonstrating the effectiveness of our question-answering pipeline in generating high-quality pairs.

With its bilingual question-answer pairs across document scenarios, XPaperQA provides a valuable benchmark for evaluating cross-lingual multimodal understanding. It enables researchers to address challenges arising from linguistic variations between visual and textual modalities and to develop potential solutions.

Evaluation Metrics

With a focus on measuring the faithfulness of answers under cross-lingual instructions, we do not care about the language of the answer as long as it is correct. We uniformly translate the answer to the source language L^{src} in the image. We use the F1 score to measure accuracy on XPaperQA.

Evaluation of LVLMs on XT-VQA

Experimental Setup We use the respective prompt set by LVLM to get its best performance and set the temperature to the default value in the model implementation. OCR tokens were provided if the model required them by default.

We benchmark 8 open-source and 2 closed-source LVLMs on XT-VQA, reporting results separately for extended datasets and the newly collected XPaperQA. Details of models and datasets are in the Appendix. Table 2 shows LVLM performance on XT-VQA, which evaluates their ability to address language inconsistencies between image text and questions—a key challenge for text-rich data like charts, tables, and documents. The benchmark analysis reveals the following findings:

Cross-lingual questions do produce a performance decline among the eight LVLMs. Although LVLMs have achieved promising accuracy conditioned on English instructions, the overall average performance of these 8 LVLMs decreased by 32.5% in Chinese and 32.6% in French. In particular, TextVQA decreased most at 34.5% in Chinese and 40.4% in French, the DocVQA follows next, decreased at 33.6% and 30.3% separately in Chinese and French. The OCRVQA has a gap of 33.0% in Chinese and 29.5% in French, while performance decreases 27.9% and 30.3% separately on ChartQA.

Involving multilingual data during training exhibits relative consistent cross-lingual performance. After calculation, compared to LLaVA-v1.6-34b 24.7% decline, Cog-VLM receives an overall 13.8% decline conditioned on Chinese queries, while their monolingual performance is close to each other. MiniCPM-V has a 14.4% decrease conditioned on French queries, which is also better than the 24.5% decline of LLaVA-v1.6.34b. We suppose it is because Cog-VLM and MiniCPM-V utilize more multilingual instruction during fine-tuning.

Mutual Information Analysis

This section analyzes the performance of LVLMs on XT-VQA from an information theory perspective. We first show how we employ mutual information to examine their cross-lingual transfer capabilities. Subsequently, we present the insights derived from our mutual information analysis.

Large Vision-Language Model Architecture

An LVLM is typically composed of a vision encoder $E(\cdot)$, a projector $f(\cdot)$, and a Large Language Model (LLM) backbone p_θ parameterized by θ. The model takes image input I with a text sequence $x = [x_1, ..., x_n]$ as the instruction together to generate a corresponding sequences $y = [y_1, ...y_n]$. The image was first encoded by vision encoder as $E_v = E(I)$ and then projected and tokenized to the text embedding space by projector $f(E(v))$ as a sequence of visual tokens $v = [v_1, ..., v_n]$. As a Markov process, the conditional probability distribution $p_\theta(y|v, x)$ can be decomposed as

$$p_\theta(y|v, x) = \prod_{i=1}^{m} p_\theta(y_i|v, x, y_{<i}). \tag{4}$$

Mutual Information in LVLM

For our cross-lingual text-rich question-answering task, we have a question Q in specific language and the outputs $Y \in \mathcal{Y}$ from LVLM, and the given image $I \in \mathcal{I}$ is tokenized as $V \in \mathcal{V}$. Given their joint distribution based on question $p(y, v|q)$, the relevance of outputs and image token v is defined as the mutual information $I(Y; V|Q)$, where V implicitly determines the distribution of Y. We want to analyze the mutual information between the outputs Y and the text-rich image tokens V conditioned on the cross-lingual question Q, formulated as

$$\begin{aligned} I(Y|V, Q) &= H(Y|Q) - H(Y|V, Q) \\ &= -\sum_{|\mathcal{Y}|} P(y|V) \log P(y|V) \\ &\quad - \sum_{|\mathcal{Y}|} P(Y|V, Q) \log P(Y|V, Q). \end{aligned} \tag{5}$$

Here, $H(Y|Q)$ represents the unconditional entropy of the output distribution which is invisible of the referenced image tokens I, while $H(Y|V, Q)$ represents the conditional entropy of the output distribution given both the image and question tokens.

Directly calculating the entropy $H(Y|V, Q)$ on the entire sentence distribution is computationally intractable due to the exponential growth of the vocabulary size $|W|$ with respect to the sequence length l. However, as a Markov process (Jelinek 1985), the probability distribution between the tokens $p(y_i|y_{i<i})$ that make Y is independent, we can decompose the entropy as:

$$\begin{aligned} H(Y|V, Q) = H(y_1, ..y_i.., y_n|V, Q) &= \sum_{i}^{|Y|} H(y_i|V, Q, y_{<i}) \\ &= \sum_{i}^{|Y|} p_\theta(y_i|V, Q, y_{<i}) \log p_\theta(y_i|V, Q, y_{<i}). \end{aligned} \tag{6}$$

Note that $H(Y|Q)$ represents the unconditional case where the LVLM cannot see the image I, meaning $V = \phi$. Since pure text input without images may cause unexpected effects on the LVLM distribution, we replace the unconditional entropy $H(Y|Q)$ with a Gaussian noise-augmented image $I_\epsilon = \epsilon + I, \epsilon \sim \mathcal{N}(\mu, \sigma)$, tokenized as V_ϵ, to ensure the stability of the final distribution. Based on the ability of heavy noise to corrupt visual information, we assume an equivalence between adding noise to the image and having no image: $H(Y|Q) = H(Y|V = \phi, Q) \approx H(Y|V_\epsilon, Q)$. This assumption follows visual contrastive decoding (Leng et al. 2024), where Gaussian noise approximates the unconditional distribution by reducing the influence of visual information and making outputs rely more on linguistic priors.

Finally, the mutual information is calculated as:

$$\begin{aligned} I(Y; V|Q) &= H(Y|V = \phi, Q) - H(Y|V, Q) \\ &= \sum_{i}^{|Y|} p_\theta(y_i|V_\epsilon, Q) \log p_\theta(y_i|V_\epsilon, Q) \\ &\quad - \sum_{i}^{|Y|} p_\theta(y_i|V, Q) \log p_\theta(y_i|V, Q). \end{aligned} \tag{7}$$

That's how we utilize mutual information to measure the extent of activation between the input image and LVLM outputs, conditioned on queries in different languages. A higher $I(Y; V|Q)$, suggests a stronger correlation between the image and the outputs under the condition of a given question, which aids in answers in faithfulness.

Mutual Information Analysis across Languages

We randomly selected 100 examples from the ChartQA dataset. After analyzing instructions in eight different languages, we arrived at two clear conclusions:

1. Entropy as a measure of uncertainty: As depicted in Figure 4, while the unconditional entropy $H(Y|V_\epsilon, Q)$ appears random, the overall distribution of conditional entropy $H(Y|V, Q)$ for correct examples is significantly lower than for incorrect examples. This reveals that the correct examples have higher certainty than incorrect examples.

2. Correlation between accuracy and mutual information: we illustrate the accuracy and mutual information of 8 different languages in Figure 3, even state-of-the-art LVLMs like Qwen-VL-Chat, which have been fine-tuned on multilingual data, display a noticeable performance disparity in cross-lingual contexts. The variation in mutual information with instructions in different languages indicates how much the visual information is activated by cross-language instructions. At last, we found a strong correlation between accuracy and mutual information, where questions in the same language as the document provide more mutual information in the answer, i.e., $I(X; Y|Q^{src}) > I(X; Y|Q^{others})$.

Methodology

In this section, we introduce **MVCL-MI** (**M**aximize **V**ision-Language **C**ross-**L**ingual **M**utual **I**nformation), to mitigate the cross-lingual performance gap on XT-VQA.

Based on the analysis, our goal is to maximize the mutual information $I^{src} = I(Y; V|Q^{src})$ between outputs and images V containing text in source language L^{src} conditioned on the question in target language L^{tgt} while retaining mu-

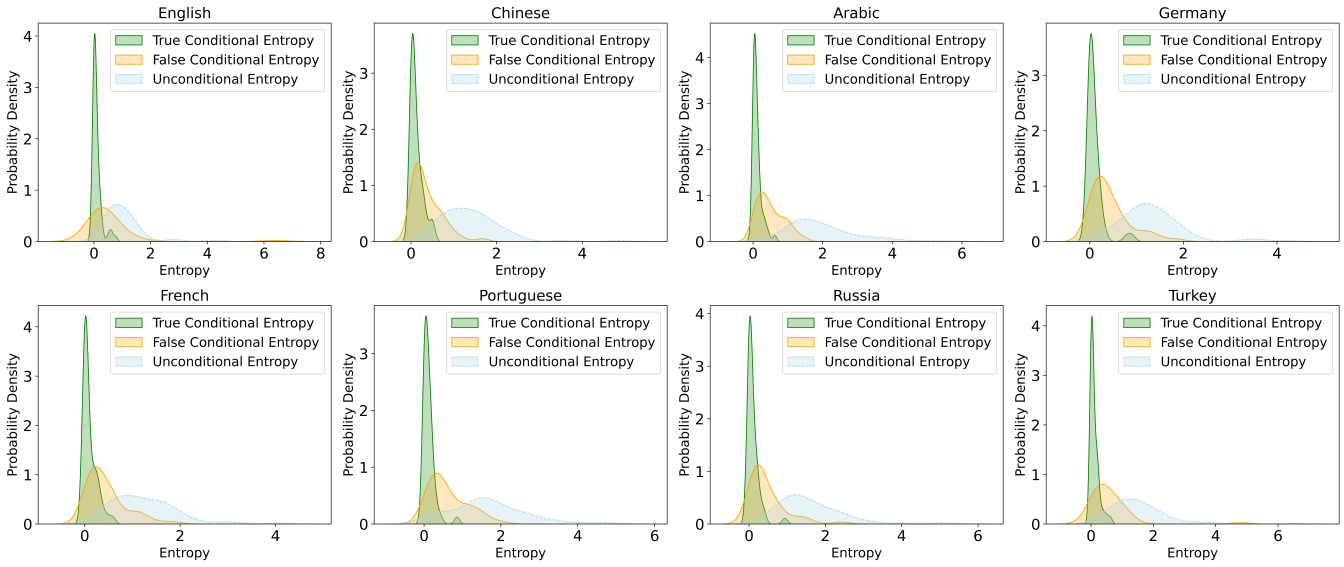

Figure 3: The entropy distribution of 100 randomly selected examples on the ChartQA dataset in 8 different languages, where the vertical axis represents probability density and the horizontal axis represents the numerical value of entropy. In all 8 languages, the mean and variance of the conditional entropy distribution for correct examples (represented in green) are significantly lower than those for incorrect examples (represented in yellow).

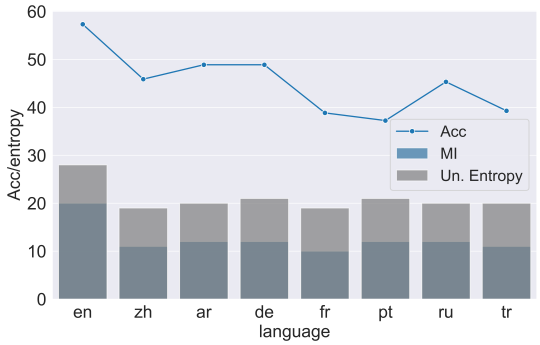

Figure 4: Statistics of accuracy and mutual information over 8 different languages on ChartQA dataset. Query in English (same in image text language) performs best, while all other languages have decreased to some extent. Reflects a correlation of accuracy and mutual information.

tual information $I^{src} = I(Y; V|Q^{src})$ conditioned on L^{src}

$$I^{src} = H(Y|V_\epsilon, Q^{src}) - H(Y|V, Q^{src})$$
$$I^{tgt} = H(Y|V_\epsilon, Q^{tgt}) - H(Y|V, Q^{tgt}) \quad (8)$$

As analyzed in Figure 4, the distribution of unconditional entropy is not affected by language, so that $H(Y|V_\epsilon, Q^{src})$ and $H(Y|V_\epsilon, Q^{tgt})$ is close to each other. To increase I^{tgt} while retaining I^{src}, we only need to minimize $H(Y|V, Q^{tgt})$.

However, directly minimizing entropy may cause the LVLM to exploit shortcuts, leading to overly sharp output logit distributions. To address this, we adopt a knowledge distillation approach by using the output logits of Q^{src} as a teacher. Specifically, we minimize the KL divergence (Kullback and Leibler 1951) between the target distribution $P_\theta^{tgt} = p_\theta(y_i|V, Q^{tgt})$ and the source distribution $P_\theta^{src} = p_\theta(y_i|V, Q^{src})$.

We add this to training objective across languages as $\mathcal{L}_{\text{KL}}$.

$$\mathcal{D}_{\text{KL}}(P_\theta^{tgt} \| P_\theta^{src}) = \sum_{i=1}^{N} p_\theta(y_i|Q^{tgt}) \cdot (\log \frac{p_\theta(y_i|Q^{tgt})}{p_\theta(y_i|Q^{src})}). \quad (9)$$

$$\mathcal{L} = \sum_{s,t \in \{\text{src,tgt}\}} \mathcal{L}_{\text{CE}}(y^{s-t}, \hat{y}^t) + \alpha \mathcal{L}_{\text{KL}}(P^{\text{src-tgt}} \| P^{\text{tgt-tgt}}). \quad (10)$$

Here, $y^{\text{src-tgt}}$ represents the output logits of LVLM $p_\theta(y^{\text{src}}|Q^{\text{tgt}})$, queried in the source language and answered in the target language, and vice versa.

$\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss for source and target language, maximizing the likelihood of the ground truth answers y_{src} and y_{tgt} given the image and question in the respective languages. D_{KL} represents the KL divergence between predicted distributions in different languages, with α as hyperparameters controlling their importance.

Experiment Results

Table 3 highlights the effectiveness of MVCL-MI by comparing the cross-lingual gap with other LVLMs. We analyze the impact of question types and language documentation on the performance gap in XPaperQA, with ablation studies confirming the necessity of our training objective.

The cross-lingual gap exists no matter the language of the source language of images. After calculating, the overall average performance gap of 8 LVLMs is 28.1% in

Model	XPaperQA-en extractive en	extractive zh	abstractive en	abstractive zh	yes-no en	yes-no zh	overall en	overall zh	XPaperQA-zh extractive zh	extractive en	abstractive zh	abstractive en	yes-no zh	yes-no en	overall zh	overall en
Open-Sourced																
LLaVA-v1.5-13b	**9.0**	5.9 -3.1	**12.3**	8.4 -3.9	**50.8**	11.3 -39.5	**14.1**	6.9 -7.2	**8.9**	3.9 -5.0	**16.2**	11.5 -4.7	**68.5**	58.5 -10.0	**27.1**	20.9 -6.2
LLaVA-v1.6-34b	**18.3**	12.0 -6.3	**23.4**	12.8 -10.6	**74.2**	57.4 -16.8	**26.8**	16.9 -9.9	**18.9**	7.5 -11.4	**24.7**	10.2 -14.5	**73.2**	66.9 -6.3	**35.2**	23.9 -11.3
InstructBLIP	**8.1**	1.8 -6.3	**15.4**	12.9 -2.5	**56.2**	50.3 -5.9	**11.3**	7.5 -3.8	**4.8**	5.2 +0.4	**9.9**	6.4 -3.5	**66.1**	52.6 -13.5	**22.6**	17.9 -4.7
mPlug-Owl2	**11.2**	9.7 -1.5	**16.3**	12.1 -4.2	**63.3**	61.7 -1.6	**17.6**	15.6 -2.0	**14.3**	2.5 -11.8	**5.5**	2.2 -3.3	**68.6**	63.9 -4.7	**24.9**	18.2 -6.7
Qwen-VL-Chat	**13.1**	9.9 -3.2	**15.1**	11.7 -3.4	**50.0**	32.7 -17.3	**16.8**	12.4 -4.4	**9.8**	6.6 -3.2	**23.8**	17.7 -6.1	**69.3**	61.6 -7.7	**30.6**	25.1 -5.5
Monkey	**20.6**	15.1 -5.5	**16.2**	6.1 -10.1	**49.2**	45.8 -3.4	**21.5**	15.4 -6.1	**17.4**	12.7 -4.7	**18.4**	14.2 -4.2	**73.1**	60.8 -12.3	**32.2**	25.7 -6.5
Cog-VLM	**16.2**	12.4 -3.8	**22.2**	14.8 -7.4	**47.5**	53.4 +5.9	**20.4**	16.8 -3.6	**21.3**	9.9 -11.4	**25.1**	16.2 -8.9	**74.5**	69.3 -5.2	**36.5**	27.7 -8.8
MiniCPM-V	**25.8**	22.3 -3.5	**14.7**	12.7 -2.0	**56.5**	37.4 -19.1	**25.3**	20.4 -4.9	**36.6**	18.5 -18.1	**34.7**	16.9 -17.8	**77.6**	75.4 -2.2	**46.1**	32.6 -13.5
ours																
MVCL-MI (8B)	**25.9**	22.5 -3.4	**16.0**	14.1 -1.9	**60.4**	52.3 -8.1	**25.9**	22.5 -3.4	**36.2**	23.9 -12.3	**37.2**	23.2 -14.0	**81.3**	79.0 -2.3	**48.2**	37.9 -10.3
Δ	↑0.1	↑0.2	↑1.3	↑1.4	↑3.9	↑15.1	↑0.6	↑1.9	↓0.4	↑5.4	↑2.5	↑6.3	↑3.7	↑3.6	↑2.1	↑5.3

Table 3: LVLMs performance on XT-VQA subset XPaperQA. The best performance over tested models is marked as Underline. Performance gap compared to L_{src} is indicated as 12.3 -3.9. Δ indicates the changes compare with MiniCPM-V.

XPaperQA-en and 24.3% in XPaperQA-zh. In the monolingual setting, LLaVA-v1.6-34b has the best performance of 26.8 in English paper, and MiniCPM performs best of 46.1 in Chinese paper. In the cross-lingual setting, MiniCPM-V performs best in both English and Chinese papers, with an accuracy of 22.5 and 32.6 respectively.

Cross-lingual gap varies in different types of questions. In English paper, the decrease on the 3 types of questions is 31.4% in abstractive, 31.0% in extractive, and 22.0% in yes-no. In Chinese paper, the average decrease in the 3 types of questions is 44.2% in extractive, 39.8% in abstractive, and 11.1% in yes-no, exhibit $Gap_{\text{ext}} > Gap_{\text{abs}} > Gap_{\text{yesno}}$. Since answering correctly to the abstractive and extractive questions demands higher comprehension of visual information than yes-no questions, it reflects that cross-lingual do affects the capabilities of LVLMs.

MVCL-MI effectively mitigates the cross-lingual gap despite the language in images. Compared to the original model, our model improved performance in cross-lingual settings, with an increase of 1.9 (↑9.3%) on XPaperQA-en and 5.3 (↑16.3%) on XPaper-zh while preserving monolingual performance. MVCL-MI narrows the performance gap from 19.4% to 13.1% (↓32.5%) in XPaper-en and 29.3% to 21.3% (↓27.3%) in XPaper-zh.

Ablation Study

We design the following ablation study to test the effectiveness of our MVCL-MI training objectives, as depicted in Tab.4. The *w/o* KL-Loss setting entails removing the KL-Loss from the training objectives, shown in F. 10. The *w/o* Cross-CE indicates we remove the cross-lingual Cross Entropy Loss from the training objectives F. 10. Removing Cross-CE lowers cross-lingual performance from 22.5 to 21.7 on English and by 4.4 on Chinese, showing the importance of cross-lingual tuning. Removing KL-Loss performs even worse than the original model. The KL divergence ensures cross-lingual predictions align with monolingual ones, maximizing mutual information between answers and input images while maintaining monolingual performance.

Method	yes-no en	yes-no zh	extractive en	extractive zh	abstractive en	abstractive zh	overall en	overall zh
English Paper								
origin	56.5	37.4	25.9	22.3	14.7	12.7	25.3	20.4
MCVL-MI	60.4	52.3	25.9	22.5	16.0	14.1	26.0 +0.7	22.5 +1.9
w/o Cross-CE	56.6	49.2	24.9	22.1	15.3	13.9	25.1 -0.2	21.7 +1.3
w/o KL-Loss	58.5	42.6	24.4	17.8	17.6	10.8	25.3 +0.0	19.7 -0.7
	zh	en	zh	en	zh	en	zh	en
Chinese Paper								
origin	77.6	75.4	35.6	18.5	34.7	16.9	46.1	32.6
MCVL-MI	81.3	79.0	36.2	23.9	37.2	23.2	48.2 +2.1	37.9 +5.3
w/o Cross-CE	77.9	76.9	35.4	19.9	36.6	17.1	46.8 +0.7	33.5 +0.9
w/o KL-Loss	73.2	70.7	34.8	17.7	31.2	19.2	43.3 -2.8	31.9 -0.7

Table 4: Ablation Study of MVCL-MI. Underline indicates the original overall performance, the best was **Bold**.

Training details We deploy our method on the advanced LVLM MiniCPM-Llama3-V. Training was done on 8 A100-sxm4-80gb for 1 epoch with the default configuration and hyperparameters. This setup is detailed in the Appendix.

Conclusion

In this paper, we investigate the cross-lingual gap in text-rich visual comprehension and propose XT-VQA, a benchmark for testing LVLMs' ability to handle language inconsistencies across modalities. From an information perspective, we identify that this gap arises from insufficient activation of visual information by cross-lingual queries. To address this, we mitigate the gap by maximizing cross-lingual mutual information. Results show that MVCL-MI enables LVLMs to effectively leverage both visual and textual information, producing accurate and language-consistent answers across languages. We believe this research advances text-rich visual comprehension and enhances LVLMs' global accessibility, fostering inclusive and cross-cultural communication.

References

Andersland, M. 2024. Amharic LLaMA and LLaVA: Multimodal LLMs for Low Resource Languages. arXiv:2403.06354.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, Y.; Zhang, H.; Zhou, Y.; Bai, X.; Wang, Y.; Zhong, M.; Yan, J.; Li, Y.; Li, J.; Zhu, M.; and Zhang, Y. 2023. Revisiting Cross-Lingual Summarization: A Corpus-based Study and A New Benchmark with Improved Annotation. arXiv:2307.04018.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Dasigi, P.; Lo, K.; Beltagy, I.; Cohan, A.; Smith, N. A.; and Gardner, M. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. arXiv:2105.03011.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Guha, N.; Nyarko, J.; Ho, D.; Ré, C.; Chilton, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.; Zambrano, D.; et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Hinck, M.; Holtermann, C.; Olson, M. L.; Schneider, F.; Yu, S.; Bhiwandiwalla, A.; Lauscher, A.; Tseng, S.; and Lal, V. 2024. Why do LLaVA Vision-Language Models Reply to Images in English? *arXiv preprint arXiv:2407.02333*.

Huang, K.; Mo, F.; Li, H.; Li, Y.; Zhang, Y.; Yi, W.; Mao, Y.; Liu, J.; Xu, Y.; Xu, J.; Nie, J.-Y.; and Liu, Y. 2024. A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers. arXiv:2405.10936.

Jelinek, F. 1985. Markov source modeling of text generation. In *The impact of processing techniques on communications*, 569–591. Springer.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Lee, K.; Joshi, M.; Turc, I.; Hu, H.; Liu, F.; Eisenschlos, J.; Khandelwal, U.; Shaw, P.; Chang, M.-W.; and Toutanova, K. 2023. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. arXiv:2210.03347.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, J.; and Jurafsky, D. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26763–26773.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Luo, C.; Shen, Y.; Zhu, Z.; Zheng, Q.; Yu, Z.; and Yao, C. 2024. LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15630–15640.

Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

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

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

Mu, Y.; Zhang, Q.; Hu, M.; Wang, W.; Ding, M.; Jin, J.; Wang, B.; Dai, J.; Qiao, Y.; and Luo, P. 2024. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36.

Nandwani, Y.; Kumar, V.; Raghu, D.; Joshi, S.; and Lastras, L. A. 2023. Pointwise Mutual Information Based Metric and Decoding Strategy for Faithful Generation in Document Grounded Dialogs. arXiv:2305.12191.

Niu, R.; Li, J.; Wang, S.; Fu, Y.; Hu, X.; Leng, X.; Kong, H.; Chang, Y.; and Wang, Q. 2024. ScreenAgent: A Vision Language Model-driven Computer Control Agent. arXiv:2402.07945.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Schneider, F.; and Sitaram, S. 2024. M5 – A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. arXiv:2407.03791.

Shin, D.; Lim, H.; Won, I.; Choi, C.; Kim, M.; Song, S.; Yoo, H.; Kim, S.; and Lim, K. 2024. X-LLaVA: Optimizing Bilingual Large Vision-Language Alignment. arXiv:2403.11399.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read. arXiv:1904.08920.

Tang, J.; Liu, Q.; Ye, Y.; Lu, J.; Wei, S.; Lin, C.; Li, W.; Mahmood, M. F. F. B.; Feng, H.; Zhao, Z.; et al. 2024. MTVQA: Benchmarking Multilingual Text-Centric Visual Question Answering. *arXiv preprint arXiv:2405.11985*.

Thakur, N.; Ni, J.; Ábrego, G. H.; Wieting, J.; Lin, J.; and Cer, D. 2024. Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval. arXiv:2311.05800.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, 1–5. IEEE.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Tschannen, M.; Mustafa, B.; and Houlsby, N. 2023. Clippo: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11006–11017.

Wan, Z.; Liu, C.; Zhang, M.; Fu, J.; Wang, B.; Cheng, S.; Ma, L.; Quilodrán-Casas, C.; and Arcucci, R. 2024. Medunic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36.

Wang, B.; Liu, Z.; Huang, X.; Jiao, F.; Ding, Y.; Aw, A.; and Chen, N. F. 2024. SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning. arXiv:2309.04766.

Wang, J.; Liang, Y.; Meng, F.; Zou, B.; Li, Z.; Qu, J.; and Zhou, J. 2023. Zero-Shot Cross-Lingual Summarization via Large Language Models. arXiv:2302.14229.

Xiao, Y.; and Wang, W. Y. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

Yang, Y.; Zhou, T.; Li, K.; Tao, D.; Li, L.; Shen, L.; He, X.; Jiang, J.; and Shi, Y. 2024. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26275–26285.

Ye, J.; Hu, A.; Xu, H.; Ye, Q.; Yan, M.; Xu, G.; Li, C.; Tian, J.; Qian, Q.; Zhang, J.; Jin, Q.; He, L.; Lin, X. A.; and Huang, F. 2023. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model. arXiv:2310.05126.

Yu, Y.-Q.; Liao, M.; Wu, J.; Liao, Y.; Zheng, X.; and Zeng, W. 2024. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *arXiv preprint arXiv:2404.09204*.