TVQACML: Benchmarking Text-Centric Visual Question Answering in Multilingual Chinese Minority Languages

Jiu Sha[1], Yu Weng[1], Mengxiao Zhu[2], Chong Feng[3], Zheng Liu[1], Jialedongzhu[1]
[1] School of Information Engineering and the Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China
[2] School of Artifical Intelligence and Computer Science, North China University of Technology
[3] School of Computer Science & Technology, Beijing Institute of Technology
Correspondence: wengyu@muc.edu.cn; zhumx@ncut.edu.cn

Abstract

Text-Centric Visual Question Answering (TEC-VQA) serves as a key benchmark for evaluating AI's ability to reason over text-rich visual scenes. However, most existing TEC-VQA datasets focus on high-resource languages and are susceptible to benchmark contamination due to overlap with pretraining corpora of large models. These limitations severely hinder progress in low-resource language scenarios and compromise the reliability of current evaluations. To address both the underrepresentation of low-resource languages and the contamination issue, we propose TVQACML, the first large-scale TEC-VQA benchmark for multilingual Chinese minority languages, constructed through a scalable, reproducible pipeline. It comprises 8,000 real-world images and 32,000 high-quality QA pairs across eight languages and 30 application scenarios. We conduct comprehensive benchmarking of open-source, closed-source, and text-centric MLLMs, revealing substantial performance gaps from human accuracy, especially in scene-text and document understanding tasks. Furthermore, instruction tuning with TVQACML yields consistent performance gains, in some cases surpassing leading closed models demonstrating the dataset's utility for model alignment. We also introduce a lightweight, extensible evaluation metric for robust multilingual, multi-format answer assessment. The code and dataset for TVQACML are available at https://github.com/Shajiu/TVQACML.

1 Introduction

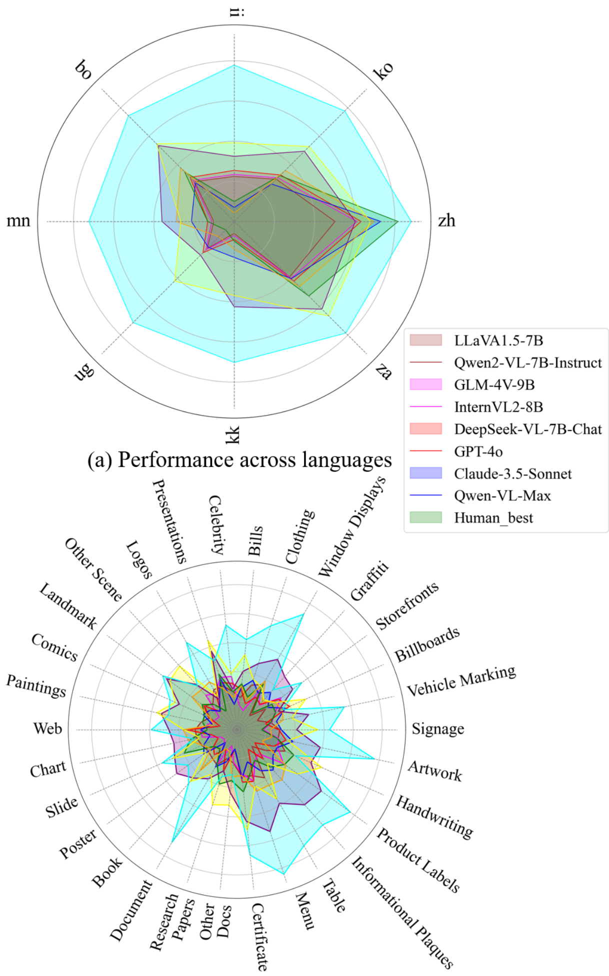
Text-Centric Visual Question Answering (TEC-VQA) (Feng et al., 2023a,b; Hu et al., 2024; Liu et al., 2024b; Tang et al., 2024a) has become a crucial benchmark for evaluating AI's ability to understand text-rich visual scenes. Unlike general VQA tasks (Biten et al., 2019; Singh et al., 2019a; Mathew et al., 2021), TEC-VQA emphasizes accurate responses based on textual content embedded within images, enabling non-specialist users to engage with complex visual information in a more accessible way. However, existing research in TEC-VQA has predominantly focused on high-resource languages such as English (Singh et al., 2019a; Mathew et al., 2021, 2022) and Chinese (Gao et al., 2015; Gan et al., 2020; Moens et al., 2021; Muresan et al., 2022), or developed regions such as Europe. This imbalance severely limits AI accessibility for low-resource language communities and hinders the equitable distribution of language technologies.

Although a few studies have attempted to expand question-answer pairs from high-resource to low-resource languages via machine translation (Changpinyo et al., 2023; Pfeiffer et al., 2022; Changpinyo et al., 2023), these approaches often suffer from severe visual-textual misalignment. Specifically, they tend to prioritize question-answer text while ignoring the actual textual content present in the image. Additionally, such methods fail to address critical issues such as nuanced semantics, contextual distortion, language bias, and question-type diversity. Compounding the issue, the open-source nature of benchmarks and the broad coverage of pretraining corpora for MLLMs have introduced benchmark contamination risks, leading to unreliable evaluation results.

Chinese minority regions, while characterized by immense linguistic diversity and spoken by millions, continue to lack sufficient support in language technologies. The absence of large-scale multilingual corpora further restricts both data availability and model development for these languages. To bridge this gap, we propose Text-Centric Visual Question Answering in Multilingual Chinese Minority Languages (TVQACML), a new benchmark specifically tailored for TEC-VQA tasks in low-resource Chinese minority languages.

We are the first to propose a full pipeline for TEC-VQA dataset construction in low-resource multilingual settings. This pipeline includes lan-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 13957–13967
November 4-9, 2025



(a) Performance across languages

(b) Performance across scenarios

Figure 1: Overall performance of MLLMs on the TVQACML benchmark.

guage selection, scenario definition, image collection, QA pair design, and quality control. It is scalable and generalizable to other low-resource languages and multimodal applications. TVQACML consists of 8,000 real-world images and 32,000 high-quality human-annotated QA pairs across eight languages: Standard Chinese (zh), Korean (ko), Sichuan Yi (ii), Tibetan (bo), Mongolian (mn), Uyghur (ug), Kazakh (kk), and Zhuang (za). It covers four major TEC-VQA scenarios—scene text, document understanding, key information extraction, and text recognition—spanning 30 application types. All data is publicly released, with standardized splits for training and evaluation.

We comprehensively benchmark representative open-source, closed-source, and text-centric MLLMs on TVQACML. Despite the relatively strong performance of closed-source models (Figure 1), all evaluated models exhibit substantial gaps compared to human performance—particularly in scene-text interpretation (e.g., signs and advertisements) and structured document understanding (e.g., tables and invoices). These findings reveal the current limitations of MLLMs in generalization and cross-lingual reasoning under text-intensive, low-resource conditions.

To further validate the quality and utility of TVQACML, we perform instruction tuning on both Chinese- and English-centric MLLMs. Experimental results show that even lightweight fine-tuning with TVQACML leads to significant performance gains, in some cases surpassing state-of-the-art closed-source models. This demonstrates the dataset's effectiveness and its strong potential for model alignment in low-resource TEC-VQA tasks.

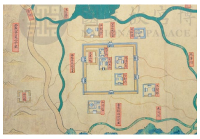
Finally, we introduce a simple yet extensible evaluation strategy that supports multilingual and multi-format answers. This metric combines character-level precision with semantic matching, enabling more robust and fair assessment across diverse model outputs.

Our main contributions are as follows:

- **A Scalable Data Construction Framework and New Benchmark**: We propose the first comprehensive data construction pipeline for TEC-VQA in low-resource multilingual settings, and introduce TVQACML, a large-scale, publicly available benchmark featuring 8,000 real-world images and 32,000 high-quality QA pairs across eight Chinese minority languages and 30 application scenarios.

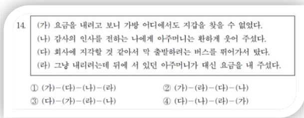
- **Comprehensive Model Benchmarking and Evaluation**: We benchmark leading open-source, closed-source, and text-centric MLLMs on TVQACML, revealing significant performance gaps compared to human annotations, particularly in scene-text and document-rich tasks, thus highlighting the current limitations of MLLMs in generalization and cross-lingual reasoning.

- **Empirical Validation of Data Utility via Instruction Tuning**: We demonstrate the effectiveness of TVQACML through instruction tuning on both Chinese- and English-centric MLLMs, achieving substantial performance improvements—sometimes surpassing top closed-source models—and propose a simple yet extensible evaluation metric to support multilingual, multi-format answer assessment.



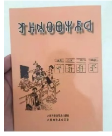
(a) Chinese (zh) example

Q: 此地图现收藏于哪里？
A: 该地图现藏于台北故宫博物院。
(Q: Where is this map currently housed?
A: This map is currently housed in the National Palace Museum in Taipei.)



(b) **Korean (ko) example**

Q: 아래 네 문장을 올바른 순서대로 배열해야 합니다.
A: ③
(Q: You need to arrange the following four sentences in the correct order.
A: ③)



(c) Sichuan Yi (ii) example

Q: [illegible]？
A: 《[illegible]》[illegible]。
(Q: What is the publisher of this book?
A: The "Yi Script Recognition Textbook" was published by Sichuan Nationalities Publishing House.)

(d) Tibetan (bo) example

Q: ད་ཐེངས་ཀྱི་གདན་ཞུའི་ལས་གནས་གང་ཡིན་ནམ།
A: ཞབས་ཕྱིའི་ལས་ཀ་བ།
(Q: What is the job position?
A: A female nurse.)

(e) Mongolian (mn) example

Q: [illegible]?
A: [illegible] 《[illegible]》

(Q: May I ask where this song is excerpted from?
A: "Sky's Colt" by Alaten Dalai and Alaten Bayar.)

(f) Uyghur (ug) example

Q: تۆۋەندىكى قايسى سائەت تاختىسى ماتورنىڭ ئايلىنىش سۈرئىتى سائىتىنى بىلدۈرىدۇ؟ a、1 رەسىم b、2 رەسىم
c、3 رەسىم d、4 رەسىم A: a
(Q: Which of the following options represents the engine tachometer? a. picture one; b. picture two; c. picture three; d. picture four A: a)

(g) Kazakh (kk) example

Q: بۇل سۋرەتتەگى قايسى ۇلت.
A: بۇل سۋرەتتەگى قازاق ۇلتى.
(Q: Which ethnic group is depicted in this image?
A: This image is of the Kazakh ethnic group.)

(h) Zhuang (za) example

Q: Raemx raeuj goengq daih?
A: C
(Q: What is the correct answer?
A: C)

Figure 2: TVQACML examples sampled from each languages. The English version in parentheses.

2 Related Work

2.1 MLLMs for Text-centric VQA

Advancements in MLLMs (penAI, 2024; Achiam et al., 2023; Yang et al., 2023; Team et al., 2023) have been transformative for VQA tasks, as evidenced by their impressive zero-shot capabilities. The ability of MLLMs to generalize, particularly after being trained on datasets focused on visual text comprehension and further refined through instruction-based fine-tuning, has greatly improved their utility in text-focused VQA contexts (Feng et al., 2023a,b; Hu et al., 2024; Liu et al., 2024b; Tang et al., 2024a). Commercially, several advanced vision-language models (VLMs), including GPT-4v (Wang et al., 2023), Gemini-Pro-V (Team et al., 2023), Qwen2-VL (Wang et al., 2024), and InternVL2 (Chen et al., 2024), have utilized publicly accessible VQA datasets related to documents to further refine text-focused VQA performance. Despite these advancements, MLLMs primarily excel in well-resourced languages like English and Chinese, leading to a performance gap for low-resource languages. This disparity is largely due to the scarcity of data and benchmarks for these languages, presenting a significant hurdle in achieving comparable results.

2.2 Multilingual Text-centric VQA Benchmarks

Progress in multilingual text-centric VQA has been driven by datasets such as GQA (Hudson and Manning, 2019), OK-VQA (Marino et al., 2019), VQAv2 (Goyal et al., 2017), and Vizwiz (Gurari et al., 2018), which benchmark visual understanding through image–question–answer annotations. To introduce greater complexity, TextVQA (Singh et al., 2019b) emphasizes OCR-based reasoning over textual content in images, while ScienceQA (Lu et al., 2022) focuses on scientific and commonsense reasoning. Recent multilingual VQA efforts include MTVQA (Tang et al., 2024c) and CVQA (Romero et al., 2024). While MTVQA centers on high-resource languages, CVQA expands coverage to global low-resource languages but pays limited attention to those within China. In contrast, our work focuses on Chinese low-resource languages such as Tibetan, Uyghur, Zhuang, and Yi, addressing their linguistic and cultural underrepresentation and offering a more equitable and context-aware VQA benchmark.

3 TVQACML Benchmark Construction

3.1 Data Collection

To ensure diversity and relevance, we collect text-rich images from both natural scenes and document contexts, as illustrated in Figure 3. Real-world images are captured through regional crowdsourcing, supplemented by web-sourced images across categories such To filter images with meaningful textual content, we apply a text retrieval model (Gómez et al., 2018) to identify those containing at least two high-confidence text instances. The retained images undergo standardized preprocessing, including multilingual OCR and language classification, and are organized by language to support subsequent annotation.

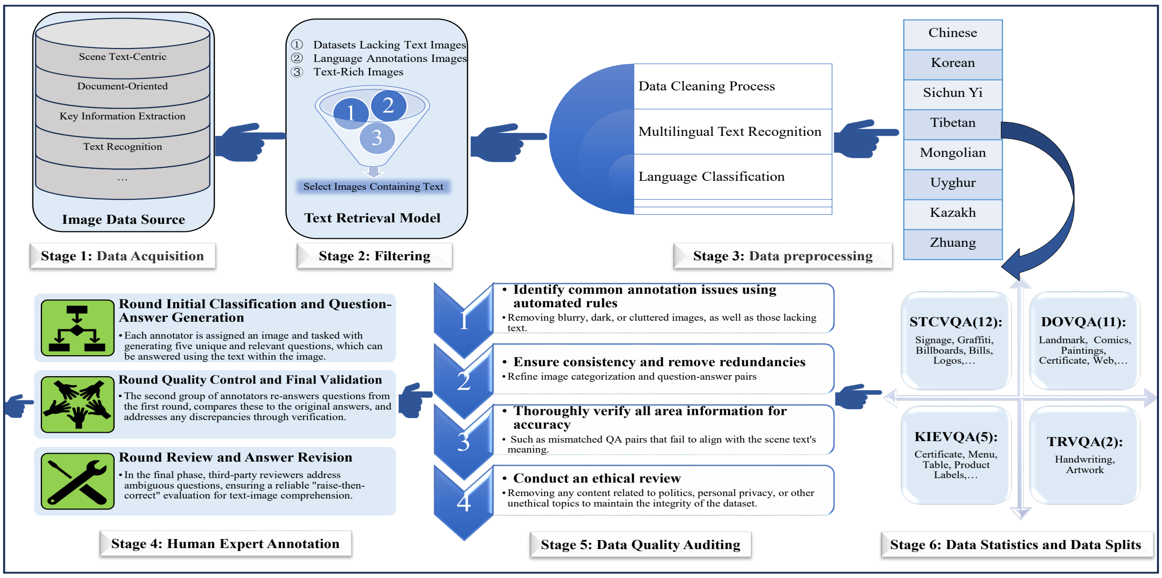


Figure 3: The construction pipeline of the TVQACML Benchmark.

3.2 Human Expert Annotation

For QA generation, we recruit native-speaking annotators with linguistic and cultural expertise in each target language. All annotators receive structured training and review sample annotations prior to the formal annotation process to ensure consistency. The annotation follows a three-stage "raise-and-verify" pipeline to ensure high-quality, diverse, and contextually grounded QA pairs.

(1) QA Generation: Annotators in the first group are assigned images and instructed to create five non-binary, unambiguous, and text-grounded questions, each with a corresponding answer. This ensures strong alignment with multilingual TEC-VQA objectives.

(2) Answer Verification: A second group independently re-answers the same questions. Discrepancies between answers are flagged for further adjudication. If two plausible answers exist, both are retained to preserve linguistic diversity.

(3) Final Review and Expert Cross-Validation: A third group of senior annotators conducts cross-validation by reviewing all ambiguous or potentially inconsistent QA pairs. These experts revise or discard low-quality items to ensure factual accuracy, linguistic clarity, and task consistency.

3.3 Quality Assurance

We implement a multi-stage quality control pipeline to ensure data integrity. First, common annotation issues (e.g., mismatched QA pairs, incomplete answers) are detected with automated rules and corrected manually by Annotators. Low-quality images—such as blurry, dark, or text-obstructed ones—are removed. Then, all annotations undergo a secondary cross-validation by senior experts to assess inter-annotator consistency and language-specific correctness. Additionally, we conduct an ethical review to remove content involving politics, privacy, or other sensitive topics. This pipeline ensures the linguistic, technical, and ethical quality of the final dataset.

3.4 Data Statistics and Data Splits

The final TVQACML dataset contains 8000 real-world images and 32000 high-quality QA pairs across eight Chinese minority languages. To ensure task coverage, each TEC-VQA scenario contains at least 100 examples, resulting in balanced coverage across 30 pre-defined abilities (Figure 4). The dataset is split into training and test sets with a 7:3 ratio to support reproducible benchmarking.

4 Experiments

Baselines. To comprehensively evaluate MLLMs' multilingual perception and comprehension capabilities, we consider three categories of models: **(1) Open-Source MLLMs:** including Qwen-VL-Chat (Bai et al., 2023), Qwen2-VL-7B-Instruct (Wang et al., 2024), LLaVA1.5-7B (Liu et al., 2024a), InternVL2-8B (Chen et al., 2024), GLM-4V-9B (GLM et al., 2024), and DeepSeek-VL-7B-Chat (Lu et al., 2024). **(2) Closed-Source MLLMs:** including GPT-4o (penAI, 2024), Claude-3.5-

Language	Qwen2-VL-7B-Instruct					GPT-4o					Qwen2-VL-7B-CML-SFT				
	chrF	CBA	Acc	vtS	Human	chrF	CBA	Acc	vtS	Human	chrF	CBA	Acc	vtS	Human
zh	45.71	52.52	42.43	41.32	42.11	69.57	69.86	51.09	50.18	50.72	47.73	41.48	34.69	34.79	35.09
ko	29.38	39.20	28.71	28.37	28.19	35.91	37.71	30.16	31.72	29.82	31.39	40.86	26.35	26.76	27.02
ii	19.43	24.98	16.45	15.31	17.23	4.22	4.09	3.62	4.79	2.75	21.45	29.54	21.10	21.30	20.36
bo	32.72	37.52	26.56	27.31	26.58	37.75	38.14	28.09	27.98	28.41	34.73	39.38	25.96	27.87	25.41
mn	2.39	7.09	2.28	1.07	2.22	23.73	26.60	10.74	9.99	11.28	4.40	11.56	4.08	5.37	4.38
ug	19.74	22.29	15.09	13.48	14.14	9.12	24.08	13.74	13.12	12.78	21.75	31.15	17.31	15.88	16.89
kk	24.08	24.99	19.33	19.78	19.52	13.74	18.37	11.20	11.48	11.42	26.09	34.59	22.32	24.12	21.98
za	37.67	43.18	32.77	31.98	33.24	45.95	46.74	38.14	38.01	37.62	39.69	50.67	35.97	35.65	35.04

Table 1: Evaluation of Multiple Models across 8 Languages with Various Metrics

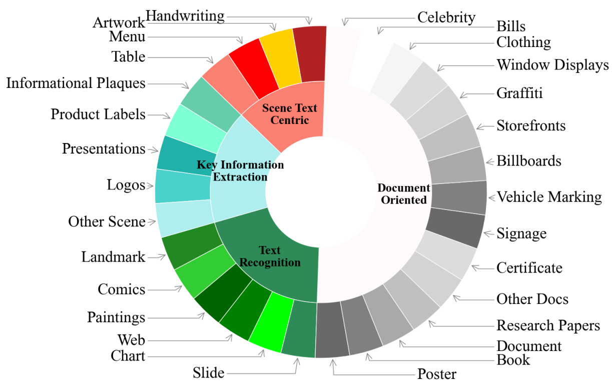


Figure 4: The TVQACML dataset covers 4 tasks and 30 subdomains. (1) **Text Recognition**: A fundamental OCR task that converts text in images into machine-readable character sequences, focusing purely on transcription without semantic understanding. (2) **Scene Text-Centric VQA**: Targets questions related to naturally occurring text in real-world scenes (e.g., signs, menus, labels), requiring both text recognition and contextual understanding. (3) **Document-Oriented VQA**: Focuses on structured or semi-structured document images (e.g., forms, invoices). Models must extract and comprehend key information based on document layout and content. (4) **Key Information Extraction VQA**: A structured VQA task aiming to extract specific key-value pairs (e.g., company, date, amount) from documents. Unlike open-ended VQA, KIE uses predefined fields and prompt-based extraction for precise matching.

Sonnet (Anthropic, 2024), GLM-4V (GLM et al., 2024), and Qwen-VL-plus/Max (Wang et al., 2024). **(3) Open-Source Text-Centric MLLMs:** including MiniCPM-V 2.6 (Yao et al., 2024), MiniCPMLlama3-V 2.5 (Yao et al., 2024) and TextSquare (Tang et al., 2024b). Additionally, we fine-tune Qwen2-VL-7B-Instruct and LLaVA1.5-7B on the TVQACML training set using the original training strategies, resulting in Qwen2-VL-7B-CML-SFT and LLaVA1.5-7B-CML-SFT. These models are specifically adapted for Chinese Minority Language TEC-VQA tasks via instruction tuning. To establish an upper-bound reference, we include a human benchmark based on five native-speaking annotators per language; we report both the best individual score (Human_best) and the average (Human_avg).

Implementation Details. All models are evaluated under zero-shot, two-shot, and five-shot settings using randomly selected in-context examples per language. Closed-source MLLMs are accessed via official APIs, while open-source models are tested through their instruct versions hosted on Hugging Face. All experiments are conducted on eight NVIDIA A800 GPUs.

Evaluation Metric. Evaluating multilingual VQA under zero-shot settings is challenging, as model responses often include explanatory or paraphrased content, making traditional metrics such as exact match or ANLS (Biten et al., 2019) less reliable. To address this, we extract language-specific spans from outputs using Unicode ranges and apply complementary metrics. We adopt chrF (Popović, 2015) for its robustness to spelling variants and morphological diversity in low-resource languages, and define a Custom Binary Accuracy (CBA) that counts a prediction as correct if it contains the ground-truth span, excluding references shorter than four characters to reduce false positives. Nonetheless, both metrics have limitations: chrF may penalize valid paraphrases, while CBA tends to overestimate correctness due to its leniency. To enhance fidelity, we further incorporate Accuracy (Acc) (the proportion of predictions that exactly match any reference) and Visual-Textual Similarity (vtS) based on CINO embeddings (Yang et al., 2022) to capture multimodal alignment. Results in Table 1 show that CBA inflates scores by 3–15% compared with human judgment, whereas Acc and vtS correlate more closely with semantic correctness and human rankings. However, vtS inherits language coverage constraints from

Models	n-shot	Languages																Avg.	
		zh		ko		ii		bo		mn		ug		kk		za			
		chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc
Open-Source MLLMs																			
Qwen-VL-Chat	0-shot	29.61	23.76	24.31	18.25	9.99	8.06	20.36	14.75	9.80	7.97	10.43	9.36	13.71	9.86	26.45	21.99	18.08	15.26
	2-shot	**33.41**	28.24	**26.98**	22.65	12.21	8.89	**23.14**	17.82	12.26	10.43	13.02	12.5	**17.08**	16.02	**29.33**	25.58	**20.93**	16.86
	5-shot	31.78	30.56	26.52	26.27	**13.01**	12.84	23.12	19.17	**12.47**	11.75	**13.21**	11.29	15.90	15.55	29.17	22.77	20.65	18.68
Qwen2-VL-7B-Instruct	0-shot	45.71	42.43	29.38	28.71	19.43	16.45	32.72	26.56	2.39	2.28	19.74	15.09	24.08	19.33	37.67	32.77	26.39	23.79
	2-shot	48.06	43.79	**33.17**	29.32	21.63	17.80	34.86	27.34	**5.61**	5.45	**23.26**	16.41	27.59	23.34	40.85	37.52	29.38	24.30
	5-shot	**49.19**	48.93	32.54	28.87	**22.89**	18.30	**36.35**	28.67	5.08	5.06	22.96	20.07	**28.15**	20.52	**40.16**	37.03	**29.67**	25.91
LLaVA1.5-7B	0-shot	37.18	28.88	24.77	22.98	14.67	10.62	24.78	18.15	2.04	1.51	16.93	16.43	17.50	15.88	29.85	27.70	21.46	18.97
	2-shot	38.60	34.14	**32.71**	25.19	**17.63**	15.66	27.28	22.18	**5.31**	3.99	17.70	16.73	21.72	21.17	30.03	27.92	23.38	19.67
	5-shot	**42.56**	32.01	30.02	28.06	16.59	13.60	**32.16**	30.27	3.95	3.77	**22.67**	19.93	**24.55**	24.42	**36.38**	34.72	**26.11**	23.35
InternVL2-8B	0-shot	42.99	30.43	26.36	21.31	6.98	5.96	27.34	22.39	21.28	16.08	18.22	13.50	17.90	16.66	40.20	35.88	25.16	21.36
	2-shot	**43.15**	37.19	26.56	22.96	**7.10**	6.16	27.47	25.48	21.45	20.86	18.26	17.62	**18.07**	17.15	**40.27**	37.09	**25.29**	21.84
	5-shot	43.09	35.65	**26.37**	24.61	7.03	6.23	**27.53**	26.08	**21.32**	18.57	**18.38**	15.95	18.04	17.10	**40.27**	38.54	25.25	22.98
GLM-4V-9B	0-shot	43.19	32.07	32.72	24.90	25.43	21.24	31.03	27.57	13.18	13.11	22.03	17.69	9.00	7.54	42.02	31.62	27.33	23.89
	2-shot	44.31	37.03	34.22	32.60	27.08	24.33	33.37	28.69	16.72	16.18	24.38	21.16	**12.77**	12.48	44.44	42.73	29.66	25.94
	5-shot	**46.28**	39.13	**35.48**	32.58	**28.19**	21.76	**33.95**	27.74	**17.15**	14.51	**25.05**	17.72	12.43	11.50	**44.76**	37.40	**30.41**	24.33
DeepSeek-VL-7B-Chat	0-shot	81.57	66.54	32.28	25.72	9.93	8.89	34.98	30.16	13.21	10.27	5.98	5.05	9.79	7.90	52.49	40.90	30.03	26.32
	2-shot	84.35	79.15	34.62	31.10	12.84	12.39	38.19	32.57	15.87	13.85	8.92	7.27	**13.19**	12.58	**56.01**	43.89	33.00	27.22
	5-shot	**84.98**	76.54	**35.40**	32.28	**12.30**	8.96	**39.74**	36.68	**15.90**	13.38	**11.04**	9.36	12.29	8.64	55.67	50.12	**33.41**	29.49
Closed-Source MLLMs																			
GLM-4V	0-shot	57.28	43.11	17.56	13.19	18.98	14.47	26.07	22.64	2.74	2.68	9.99	8.08	1.26	0.9	38.80	30.78	21.59	17.84
	2-shot	**60.21**	46.72	19.19	16.71	**21.57**	15.27	28.09	24.42	4.35	3.35	**12.50**	11.96	3.22	2.59	40.31	32.23	23.68	19.88
	5-shot	60.11	49.01	**20.38**	18.86	20.13	15.17	**28.58**	27.56	**4.46**	3.15	11.38	9.4	**4.10**	3.95	**40.74**	32.63	**23.73**	18.38
Qwen-VL-plus	0-shot	54.00	40.93	33.53	25.79	20.63	17.87	29.85	22.47	14.94	10.62	21.17	19.84	24.32	19.15	37.10	31.16	29.44	25.52
	2-shot	55.17	46.10	34.79	27.86	21.83	19.55	31.80	25.45	16.28	15.98	23.13	19.85	26.04	20.09	39.01	31.52	31.01	25.61
	5-shot	**56.84**	54.63	**36.36**	29.26	**23.71**	21.77	**32.06**	29.03	**17.39**	16.54	**23.84**	21.44	**26.46**	19.97	**39.59**	34.38	**32.03**	26.53
GPT-4o	0-shot	69.57	51.09	35.91	30.16	4.22	3.62	37.75	28.09	28.90	23.73	10.74	9.12	13.74	11.2	45.95	38.14	30.85	24.98
	2-shot	71.07	62.39	37.18	36.63	5.83	4.68	39.51	39.26	30.67	28.22	12.08	9.67	15.17	15.11	47.52	42.3	32.38	30.27
	5-shot	**72.11**	61.7	**38.27**	33.3	**6.85**	6.35	**39.97**	36.82	**31.00**	25.55	**13.49**	10.4	**16.27**	13.83	**48.46**	46.75	**33.30**	28.26
Claude-3.5-Sonnet	0-shot	60.26	45.36	49.46	41.16	32.44	25.78	53.48	52.2	35.95	32.51	23.73	19.49	42.48	33.29	61.77	45.17	44.95	39.42
	2-shot	61.42	57.3	50.79	48.51	34.00	32.2	55.34	54.47	37.40	32.68	24.88	23.47	44.33	36.95	63.53	53.37	46.46	41.13
	5-shot	**62.70**	62.68	**51.85**	46.79	**34.56**	27.86	**56.39**	54.87	**38.36**	37.39	**26.24**	23.63	**45.55**	36.86	**64.57**	47.85	**47.53**	40.93
Qwen-VL-Max	0-shot	67.84	50.26	52.75	48.75	39.17	28.26	54.67	50.32	26.57	21.44	41.98	32.73	36.84	35.55	66.58	55.27	48.30	40.32
	2-shot	68.04	67.81	52.84	41.52	39.19	28.23	**54.78**	52.09	**26.72**	20.00	42.03	35.61	36.84	27.57	66.67	62.69	48.39	41.94
	5-shot	**67.94**	47.66	**52.92**	51.70	**39.44**	38.36	54.77	44.89	26.70	24.78	**42.14**	30.41	**36.88**	36.62	**66.85**	50.96	**48.45**	40.67
Open-Source Text-Centric MLLMs																			
MiniCPM-V 2.6	0-shot	62.16	59.58	13.43	10.74	10.78	10.77	13.58	10.9	3.85	3.29	0.00	0.0	0.00	0.0	40.00	33.91	17.98	16.15
	2-shot	**62.18**	48.92	**13.45**	12.03	10.80	10.38	13.59	11.71	3.86	3.66	0.01	0.01	0.01	0.01	**40.02**	33.78	17.99	15.06
	5-shot	62.17	51.61	**13.45**	9.79	**10.80**	10.13	**13.59**	13.12	**3.87**	3.75	**0.02**	0.02	**0.02**	0.01	40.01	30.79	**17.99**	14.90
MiniCPMLlama3-V 2.5	0-shot	50.00	45.98	26.67	22.04	15.33	14.03	17.11	14.73	16.05	13.74	13.33	13.05	9.78	8.54	40.54	32.5	23.60	20.58
	2-shot	50.01	35.95	26.68	26.42	15.35	13.95	17.12	13.81	16.07	15.17	13.34	11.66	**9.79**	9.12	40.55	36.49	23.61	20.32
	5-shot	**50.01**	49.92	**26.69**	20.89	**15.36**	15.09	**17.12**	16.13	**16.08**	13.20	**13.35**	12.04	**9.79**	7.13	**40.56**	28.64	**23.62**	20.38
TextSquare	0-shot	40.13	28.55	22.61	16.45	11.16	8.56	12.91	10.32	11.29	9.58	10.45	9.3	8.93	6.63	30.55	25.65	19.55	15.57
	2-shot	**43.56**	30.87	26.21	18.44	11.38	8.82	**14.57**	11.49	**15.39**	12.9	**12.12**	9.31	**9.46**	7.67	**39.24**	34.15	**21.07**	16.39
	5-shot	**42.63**	36.61	**26.24**	24.41	**11.73**	9.01	13.18	12.53	13.46	15.1	11.50	11.45	9.30	7.73	34.38	27.18	19.69	17.13
Instruction Tuning Models																			
LLaVA1.5-7B-CML-SFT	0-shot	39.84	36.13	26.47	20.55	15.09	14.92	29.13	22.19	3.60	3.36	19.70	15.65	19.16	14.41	27.79	25.57	23.09	19.23
	2-shot	**47.10**	44.89	29.18	22.88	20.22	18.70	**32.76**	29.34	**6.88**	5.99	20.06	16.71	22.95	20.97	**40.56**	30.04	**27.12**	23.56
	5-shot	46.46	43.93	**27.95**	20.63	**21.88**	20.30	27.97	26.41	6.26	5.92	**20.94**	17.22	**25.67**	22.12	33.29	31.81	26.15	23.54
Qwen2-VL-7B-CML-SFT	0-shot	47.73	34.69	31.39	26.35	21.45	21.10	34.73	25.96	4.40	4.08	21.75	17.31	26.09	22.32	39.69	35.97	28.40	24.28
	2-shot	51.20	49.45	33.66	27.54	24.24	21.88	37.08	30.44	**7.85**	5.78	25.35	17.94	29.27	24.30	**43.05**	42.98	31.46	26.73
	5-shot	**52.19**	43.48	**35.49**	30.18	**26.12**	25.49	**38.42**	29.01	7.61	6.65	**26.82**	23.74	**29.57**	24.81	43.03	38.75	**32.41**	27.76
Human Performance																			
Human	avg	88.14	73.74	67.82	61.37	57.88	42.81	74.48	60.92	62.34	43.82	51.20	48.8	60.22	48.0	78.61	66.15	67.59	55.7
	best	92.04	67.6	70.11	59.71	60.73	44.03	78.41	75.11	66.11	60.28	53.37	49.55	62.55	49.41	81.80	78.0	70.64	60.46

Table 2: Overall results of different MLLMs on TVQACML benchmark. Scores are marked with **bold** for the best, underline for the second-best, and red for the lowest.

CINO, reducing robustness in low-resource settings. Therefore, we adopt Acc as the primary metric and chrF as the secondary reference, which together provide a balanced and reliable evaluation protocol for multilingual VQA in low-resource and culturally diverse scenarios.

Contamination Check and Verification. To confirm the dataset's integrity, we conducted two key contamination checks: (1) Vocabulary Analysis: Except for Chinese, the remaining seven minority languages do not appear in the vocabularies of major base models, minimizing the chance of prior exposure. (2) Attribution Testing: Mainstream models (e.g., GPT-4, Qwen) failed to reproduce or complete samples from TVQACML, even under partial prompts. These results support that TVQACML is contamination-free and suitable for benchmarking multilingual TEC-VQA models.

4.1 Evaluation Results

Performance Gap: MLLMs vs. Humans. As shown in Table 2, there remains a substantial gap between MLLMs and human performance on TVQACML. In the zero-shot setting, human annotators achieve an average chrF score of 67.59% (best: 70.64%). In contrast, the best closed-source model (Qwen-VL-Max) reaches 48.30%, while the best open-source model (DeepSeek-VL-7B-Chat) achieves only 30.03%. These results underscore

Models	Languages																Avg.	
	zh		ko		ii		bo		mn		ug		kk		za			
	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc
	Scene Text-Centric VQA Task																	
Qwen2-VL-7B-Instruct	**36.21**	33.21	31.85	26.55	14.35	14.2	27.79	19.96	3.32	2.66	21.66	16.15	24.37	22.41	33.86	31.70	22.20	18.33
GLM-4V-9B	**44.39**	32.51	31.91	28.6	26.17	18.37	31.69	29.98	9.59	6.79	24.51	17.77	8.41	6.67	41.63	38.18	27.29	23.92
LLaVA1.5-7B	**32.88**	31.29	28.97	24.32	12.39	11.38	24.28	17.41	3.11	2.61	18.36	13.62	21.59	18.14	27.08	28.71	20.17	15.76
InternVL2-8B	**46.81**	37.85	26.98	19.97	8.28	8.16	26.69	26.09	18.68	14.73	17.42	14.44	21.74	15.55	33.43	20.93	25.00	20.90
DeepSeek-VL-7B-Chat	**79.21**	78.06	27.57	22.36	9.83	7.26	32.77	31.7	13.72	9.88	6.43	4.85	9.63	8.23	51.63	46.18	28.85	26.06
GPT-4o	**56.99**	52.56	36.11	26.14	4.29	3.40	45.15	44.9	31.99	26.68	9.18	9.16	13.54	11.69	54.54	45.95	31.47	27.56
Claude-3.5-Sonnet	57.23	47.33	42.54	29.92	26.03	18.4	51.95	49.36	36.3	32.02	22.02	21.45	34.11	31.17	**60.41**	53.55	41.32	35.40
Qwen-VL-Max	**72.06**	55.0	51.67	40.62	32.92	27.34	55.44	39.04	27.07	21.80	45.23	34.19	47.12	42.8	69.94	53.32	50.18	39.26
LLaVA1.5-7B-CML-SFT	36.18	32.15	29.01	24.94	13.0	12.03	28.02	19.64	3.24	2.97	20.6	15.32	25.08	20.55	31.82	33.31	21.96	16.00
Qwen2-VL-7B-CML-SFT	**41.19**	40.37	37.4	30.93	16.52	16.21	31.78	23.88	4.04	3.31	24.50	18.60	27.67	25.01	38.56	38.18	27.71	22.51
Human_best	**90.58**	86.05	71.65	50.17	49.68	39.79	70.42	52.55	63.36	57.99	47.92	40.04	59.49	56.13	76.87	73.79	66.25	57.06
	Document-Oriented VQA Task																	
Qwen2-VL-7B-Instruct	**46.55**	33.0	23.77	19.99	27.66	20.17	25.33	20.75	4.03	3.22	18.55	18.48	24.05	18.10	30.14	23.66	25.01	19.67
GLM-4V-9B	29.08	27.64	28.32	21.1	21.1	20.77	26.74	22.93	10.16	7.64	20.50	17.89	9.95	8.32	**38.45**	35.39	23.04	20.21
LLaVA1.5-7B	**39.34**	35.55	21.2	19.36	26.45	21.47	23.22	20.69	3.77	2.92	17.4	13.59	21.73	15.27	28.82	27.19	22.74	19.50
InternVL2-8B	32.05	30.39	21.11	15.06	6.39	5.43	22.31	18.69	19.98	14.84	18.76	14.04	11.90	10.05	**38.19**	29.5	21.34	17.25
DeepSeek-VL-7B-Chat	**81.52**	66.74	30.39	25.62	8.03	7.0	33.20	23.51	10.89	7.79	4.49	4.38	8.59	7.86	35.54	32.20	26.58	21.89
GPT-4o	**53.28**	37.81	36.89	35.59	3.25	2.52	26.16	18.82	25.85	23.66	9.26	8.13	14.26	10.82	37.35	35.23	25.79	21.57
Claude-3.5-Sonnet	**50.65**	38.49	45.58	36.52	30.09	25.67	46.3	37.77	41.16	33.98	23.96	23.58	37.88	37.44	47.56	47.22	40.40	35.08
Qwen-VL-Max	58.02	54.78	49.35	35.15	36.4	31.38	54.04	49.96	26.08	24.90	26.13	25.11	38.02	31.31	**59.65**	49.99	43.46	37.82
LLaVA1.5-7B-CML-SFT	**46.51**	46.38	24.58	19.89	29.43	21.17	25.78	21.69	4.34	4.05	19.60	19.59	23.93	17.04	32.61	23.23	25.85	21.63
Qwen2-VL-7B-CML-SFT	**47.95**	42.8	28.15	20.31	28.09	20.37	29.56	25.96	4.32	3.74	19.11	15.10	25.72	23.18	31.78	29.71	26.84	22.65
Human_best	**74.06**	52.95	63.27	45.74	54.05	44.97	64.45	52.07	62.79	58.14	43.87	43.1	53.36	53.33	69.03	53.36	60.61	50.46
	Key Information Extraction Task																	
Qwen2-VL-7B-Instruct	**46.49**	43.52	26.3	22.51	11.21	8.92	37.07	26.02	4.14	3.31	24.25	17.66	19.63	14.93	36.03	26.34	25.64	20.40
GLM-4V-9B	45.37	36.57	33.2	29.84	28.17	20.70	34.59	29.45	20.26	18.96	12.21	12.21	7.62	6.62	**53.24**	51.66	29.33	25.75
LLaVA1.5-7B	**39.83**	34.79	22.69	21.77	10.43	8.69	25.37	24.60	3.70	3.65	20.82	15.64	18.56	18.22	31.51	23.32	21.61	18.84
InternVL2-8B	41.37	33.26	30.96	25.18	6.61	5.94	25.41	21.46	24.50	20.85	16.77	14.71	19.34	19.13	**52.81**	38.06	27.22	22.32
DeepSeek-VL-7B-Chat	**79.79**	71.75	36.12	34.18	10.18	8.52	32.42	32.27	14.16	10.10	6.88	5.38	11.24	10.34	64.08	63.38	31.86	29.49
GPT-4o	**84.74**	66.36	40.38	38.63	4.99	4.23	38.46	34.82	27.68	23.59	8.74	7.28	18.97	18.16	47.93	36.49	33.99	28.70
Claude-3.5-Sonnet	68.05	61.83	65.69	61.71	30.73	23.83	58.87	50.33	34.07	29.81	20.04	15.94	55.5	44.63	**72.23**	65.88	50.65	44.24
Qwen-VL-Max	**64.52**	46.10	54.02	51.26	41.68	32.58	58.09	49.82	27.51	23.53	50.64	46.42	28.05	19.83	60.74	55.97	48.16	40.69
LLaVA1.5-7B-CML-SFT	**47.66**	43.74	22.81	18.14	10.85	8.68	28.25	21.59	4.05	3.97	22.29	21.89	21.05	18.26	36.45	32.35	24.18	21.08
Qwen2-VL-7B-CML-SFT	**52.30**	41.21	31.27	28.49	13.78	10.75	41.62	33.99	5.08	4.50	27.6	21.24	24.28	17.2	42.26	31.03	29.77	23.55
Human_best	**92.54**	77.65	65.01	47.78	60.75	50.93	88.12	66.12	63.49	53.51	53.85	37.91	58.75	44.55	88.39	63.56	71.36	55.25
	Text Recognition Task																	
Qwen2-VL-7B-Instruct	**35.88**	35.05	21.25	18.74	19.59	17.04	28.1	26.47	3.10	2.82	12.07	10.90	19.46	15.11	34.67	33.40	21.77	19.94
GLM-4V-9B	**49.24**	42.03	37.43	35.33	26.26	20.17	31.09	30.68	12.71	11.50	30.90	27.63	10.02	8.40	39.42	34.82	29.63	26.32
LLaVA1.5-7B	**34.23**	31.69	19.26	18.69	16.69	13.91	23.31	21.16	2.89	2.51	10.90	9.32	16.97	15.38	29.80	28.43	19.26	17.64
InternVL2-8B	40.58	29.53	26.4	23.02	6.65	5.73	34.97	34.25	21.94	19.31	19.92	19.87	18.63	17.27	**47.52**	38.18	27.08	23.40
DeepSeek-VL-7B-Chat	**85.77**	84.89	35.01	32.45	11.69	8.77	41.53	30.15	14.06	13.91	6.11	4.48	9.69	8.87	8.70	56.91	32.82	30.05
GPT-4o	**83.28**	78.66	30.25	27.81	4.35	3.46	41.22	34.46	30.09	26.04	15.78	15.06	8.20	8.11	43.98	42.01	32.14	29.45
Claude-3.5-Sonnet	65.11	62.52	44.02	34.8	42.90	34.25	56.79	55.43	32.25	24.49	28.92	20.65	42.43	34.49	**66.86**	55.50	47.41	40.27
Qwen-VL-Max	**76.78**	73.09	55.96	52.32	45.69	41.82	51.12	46.87	25.62	21.07	45.91	34.89	34.15	31.69	75.98	66.44	51.4	46.02
LLaVA1.5-7B-CML-SFT	**38.95**	35.66	22.98	16.4	21.51	18.72	28.10	27.22	3.28	2.53	12.49	9.71	21.18	19.22	36.09	31.42	23.07	20.11
Qwen2-VL-7B-CML-SFT	**49.48**	44.72	28.73	28.21	27.41	21.37	35.97	34.90	4.16	3.74	15.78	14.71	26.69	25.24	46.16	41.45	29.30	26.79
Human_best	**95.38**	87.52	71.34	64.49	67.03	51.31	74.95	69.45	59.71	42.08	59.15	44.11	69.29	61.89	80.13	71.55	72.12	61.55

Table 3: Overall results of different models on different domains (under the zero-shot setting).

Models	Languages																Avg.		$\pm\triangledown$
	zh		ko		ii		bo		mn		ug		kk		za				
	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	chrF	Acc	
	Text Recognition Task (PS)																		
Qwen2-VL-7B-Instruct	**41.01**	32.32	22.22	17.85	20.27	14.76	30.19	29.56	3.17	2.83	12.11	9.76	19.61	16.07	36.46	34.18	23.13	19.67	
GLM-4V-9B	**51.45**	47.74	39.89	36.32	26.65	26.09	29.97	22.1	12.2	8.67	28.49	22.53	10.21	9.75	41.19	38.41	30.01	26.45	
LLaVA1.5-7B	**35.95**	32.21	20.16	17.6	20.07	18.43	27.52	24.69	2.73	2.35	10.72	9.26	18.72	13.93	34.45	31.89	21.29	18.8	
InternVL2-8B	39.01	36.44	28.41	26.35	6.74	6.35	34.89	29.52	22.08	21.94	19.29	16.21	18.75	15.90	**51.09**	38.47	27.53	23.90	
DeepSeek-VL-7B-Chat	**93.57**	67.35	36.35	35.60	10.77	8.92	40.09	29.49	13.17	11.53	5.87	4.19	9.31	6.58	62.77	59.51	33.99	27.90	
GPT-4o	**77.25**	68.2	31.32	23.84	4.60	3.28	44.79	41.94	27.80	25.03	16.03	11.95	8.84	7.55	47.68	47.32	32.29	28.64	
Claude-3.5-Sonnet	58.89	56.6	41.51	32.08	42.86	34.87	58.92	51.59	29.88	21.66	29.16	28.14	38.68	32.23	**68.49**	51.12	46.05	38.54	
Qwen-VL-Max	74.13	57.08	53.47	48.91	47.66	43.83	55.04	53.98	28.16	26.83	45.88	43.05	34.9	29.33	**80.83**	62.48	52.51	45.69	
LLaVA1.5-7B-CML-SFT	**46.20**	32.35	22.45	21.80	21.79	15.98	33.36	31.53	3.49	3.18	12.95	10.32	21.43	17.13	40.23	34.01	25.24	20.79	
Qwen2-VL-7B-CML-SFT	**51.96**	43.46	27.86	20.16	25.84	22.86	38.33	37.85	3.97	3.34	15.4	15.07	24.79	24.52	45.83	40.47	29.25	25.97	
	Text Recognition Task (IS)																		
Qwen2-VL-7B-Instruct	**40.99**	32.31	22.2	17.83	20.25	14.74	30.18	29.54	3.16	2.82	12.09	9.75	19.59	16.05	36.45	34.16	23.11	19.65	0.02↓
GLM-4V-9B	**51.44**	47.72	39.87	36.31	26.63	26.07	29.96	22.08	12.18	8.65	28.47	22.52	10.19	9.73	41.18	38.39	29.99	26.43	0.02↓
LLaVA1.5-7B	**35.93**	32.2	20.15	17.59	20.05	18.41	27.50	24.68	2.72	2.33	10.70	9.25	18.7	13.91	34.43	31.87	21.27	18.78	0.02↓
InternVL2-8B	38.99	36.42	28.39	26.34	6.73	6.33	34.88	29.50	22.07	21.93	19.28	16.20	18.73	15.89	**51.07**	38.46	27.52	23.88	0.02↓
DeepSeek-VL-7B-Chat	**93.55**	67.34	36.33	35.59	10.75	8.91	40.07	29.48	13.16	11.51	5.85	4.17	9.3	6.57	62.75	59.49	33.97	27.88	0.02↓
GPT-4o	**77.23**	68.19	31.31	23.83	4.58	3.26	44.78	41.93	27.79	25.02	16.01	11.94	8.82	7.54	47.67	47.3	32.27	28.63	0.01↓
Claude-3.5-Sonnet	58.87	56.59	41.5	32.06	42.85	34.86	58.9	51.58	29.87	21.65	29.15	28.13	38.66	32.22	**68.47**	51.11	46.03	38.53	0.01↓
Qwen-VL-Max	74.12	57.07	53.46	48.89	47.65	43.81	55.02	53.96	28.14	26.82	45.87	43.03	34.88	29.31	**80.81**	62.47	52.49	45.67	0.02↓
LLaVA1.5-7B-CML-SFT	**44.52**	29.92	20.61	19.94	19.86	13.72	31.03	29.82	1.72	1.02	11.34	8.56	19.07	14.93	38.58	31.97	23.34	18.73	2.06↓
Qwen2-VL-7B-CML-SFT	**49.94**	41.41	25.87	18.33	23.95	20.78	36.73	36.23	2.41	1.07	13.38	12.72	22.91	22.1	43.43	38.45	27.33	23.89	2.08↓

Table 4: Overall results of different MLLMs on text recognition task. $\pm\triangledown$: the score gap between IS and PS sets.

the challenges MLLMs face in multilingual text-centric VQA, particularly in understanding visual semantics and producing grounded answers across diverse languages.

Language-wise Analysis. Figure 1 reveals that model performance varies significantly across languages. High-resource languages such as Chinese and Zhuang (which shares similar lexical and grammatical traits with Chinese and English) yield better results. Tibetan and Korean also perform moderately well due to their similar linguistic features with high-resource languages. In contrast, Kazakh, Uyghur, Mongolian, and Sichuan Yi perform poorly, attributed to factors such as phonetic alphabets, right-to-left scripts, and limited representation in pretraining corpora. These findings emphasize the persistent disparity between high- and low-resource languages in current MLLMs.

Model-wise Analysis. As shown in Table 2, most models perform best on Chinese tasks, with Claude-3.5-Sonnet slightly outperforming on Zhuang, possibly due to cross-cultural image descriptions in its training data. Across models, closed-source MLLMs generally outperform open-source ones (except DeepSeek-VL-7B-Chat). Qwen-VL-Max achieves the highest average chrF score (48.30%) among all models. Among open-source models, DeepSeek-VL-7B-Chat leads, especially on Chinese (81.57% chrF). However, text-centric models such as MiniCPM-V 2.6 offer only marginal gains, primarily benefiting high-resource scenarios.

Effectiveness of Instruction Tuning. Fine-tuning on TVQACML significantly enhances model performance. As shown by LLaVA1.5-7B-CML-SFT and Qwen2-VL-7B-CML-SFT, instruction tuning yields consistent improvements across all languages and tasks. Compared to their base counterparts, both models show notable gains in chrF and Acc, particularly in low-resource languages such as Kazakh, Uyghur, and Mongolian. These improvements are evident across all evaluation settings (0-shot, 2-shot, 5-shot), demonstrating the dataset's well-structured supervision and its adaptability to few-shot learning. Moreover, fine-tuned models exhibit competitive or near-human performance in some chrF cases, highlighting the benchmark's effectiveness not only as a training resource but also as a rigorous evaluation standard. The improvements span multiple task types—text recognition, scene text VQA, document VQA, and key information extraction—validating the dataset's generality and its capacity to enhance multilingual visual text understanding across architectures (e.g., LLaVA and Qwen series).

4.2 Ablation Study

To further investigate the capabilities and limitations of state-of-the-art MLLMs in low-resource multilingual TEC-VQA tasks, we conduct an in-depth ablation analysis across four task types. Detailed results are presented in Table 3.

Scene Text-Centric VQA. Despite recent progress in MLLMs, our evaluation exposes a critical gap between semantic localization and visual-linguistic generalization. Models such as DeepSeek-VL-7B-Chat perform well on high-resource languages like zh, yet their accuracy drops drastically on structurally and scriptually distinct minority languages—for instance, falling to just 6.43% on ug. Interestingly, Claude-3.5-Sonnet achieves relatively strong results on za, hinting that some closed-source models may implicitly benefit from broader multilingual exposure or more effective cross-lingual representation alignment. These disparities go beyond simple language imbalance, highlighting a deeper vulnerability: the inability of current MLLMs to generalize in scene-text VQA when deprived of dominant-script priors.

Document-Oriented VQA. Performance on document-based tasks is significantly lower than scene text tasks, reflecting the added difficulty of parsing structured layouts and hierarchical content. The challenge is amplified by the syntactic and visual complexity of documents in low-resource languages such as Kazakh (kk), Uyghur (ug), and Mongolian (mn). These languages also introduce script directionality challenges (e.g., right-to-left (kk, ug) and top-to-bottom (mn)), which current models are ill-equipped to handle.

Key Information Extraction (KIE). Results on KIE tasks reveal performance imbalance across languages. While models like GPT-4o achieve reasonable score on zh and za, they struggle on languages such as ii, possibly due to the complex structure and varying length of ii texts, which require MLLMs to capture fine-grained details accurately.

Text Recognition. Text recognition remains challenging, particularly without strong semantic cues. To investigate this, we design two test settings: Positive Sequence (PS) with natural text order and Inverse Sequence (IS) with shuffled text. Interestingly, models without instruction fine-tuning perform consistently on IS, while instruction-tuned models show a marked drop. This suggests two points: (1) Pre-tuned models lack semantic priors and are unaffected by disrupted context, supporting the absence of data leakage; (2) Fine-tuned models acquire semantic understanding from our dataset, as their performance degrades when such context is removed. These results validate the semantic richness and evaluative effectiveness of our dataset for low-resource language modeling.

Error Typology and Analysis. We conducted a qualitative error analysis across languages and task types in the TVQACML benchmark and identified five major error types, with **(1) text recognition failures** being the most dominant, e.g. models often omit entire words in Uyghur and Mongolian street signs. **(2) Cross-modal misalignment** is also common, where models incorrectly extract information from irrelevant regions. In **(3) low-resource scripts** like Tibetan and Kazakh, models frequently fail to handle special glyphs or non-standard writing directions. Additionally, in **(4) document-oriented VQA**, models misinterpret table structures, leading to mismatched labels and values. Finally, **(5) reasoning failures** arise in multi-step questions requiring comparison or sequencing, such as identifying the earlier of two dates. These issues highlight the need for enhanced OCR accuracy, cross-modal grounding, multilingual robustness, and reasoning capabilities in future MLLMs.

A deeper qualitative review indicates that these errors cannot be fully explained by high-level factors alone. Smaller models often exhibit vision–text decoupling, misreferencing objects or attributes across modalities. In low-resource languages, hallucination rates increase, with models fabricating entities or properties. Moreover, cross-linguistic variations—such as multi-script usage, diacritics, agglutination, and complex morphology—reduce tokenization fidelity and hinder effective transfer from pretrained bases. Instruction-tuning further amplifies these issues: differences in knowledge density and task difficulty across languages lead to inconsistent adherence to answer formats, resulting in verbosity or multiple, drifting answers. Overall, bad cases cluster into four recurring patterns: misgrounded references, hallucinated facts, script/orthography-induced semantic drift, and instruction non-compliance. These patterns are particularly prominent in typologically distant, multi-script, and low-coverage languages, as well as in smaller models.

5 Conclusion

We presents TVQACML, a low-resource multilingual text-Centric VQA dataset. It includes 8 languages, 4 tasks, and 30 scenarios, with each question offering multiple plausible answers. TVQACML is the first multilingual VQA benchmark fully relying on human annotations, tailored for text-centric scenarios. Experimental results show that current models still have significant room for improvement in low-resource multilingual text-centric scenarios, with performance gaps compared to human experts. TVQACML will provide valuable evaluation tools and contribute to the development of expert-level AGI.

Limitation

The current version of the TVQACML dataset, while dialectally diverse, has limitations in language coverage. Despite encompassing many languages, it lacks inclusivity, omitting numerous lesser-spoken ones. Currently, experiments cover only some domestic ethnic languages, not yet others. In the future, we will integrate all ethnic languages in China with a written form, ensuring broader representation across the linguistic spectrum. Additionally, the dataset now provides a single canonical response per question, which may not fully capture the range of answers for different expressions of the same semantics. Recognizing this, future versions will include multiple plausible answers to reflect varied perspectives.

Acknowledgments

This study was funded in part by the National Key Research and Development Program of Hainan Province, China under Grant ZDYF2024 (LALH) 005, in part by the Beijing Municipal Science & Technology Commission under Grant Z231100001723002, and in part by the National Key R&D Program of China under grant 2022YFF0902500, in part by the 2024 National Social Science Fund of China (NSSFC) – Special Project of Unpopular Extinctive Subjects under grant 24VJXG063.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V. Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. Maxm: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2667–2682. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2023a. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *CoRR*, abs/2311.11810.

Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. 2023b. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *CoRR*, abs/2308.11592.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2296–2304.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Lluís Gómez, Andrés Mafla, Marçal Rusinol, and Dimosthenis Karatzas. 2018. Single shot scene text retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 700–715.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *CoRR*, abs/2403.12895.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024b. Textmonkey: An ocr-free large multimodal model for understanding document. *CoRR*, abs/2403.04473.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors. 2021. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics.

Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. 2022. Findings of the association for computational linguistics: Acl 2022. In *Findings of the Association for Computational Linguistics: ACL 2022*.

penAI. 2024. Gpt-4o main page. *https://openai.com/index/hello-gpt-4o*.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulic, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2497–2511. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019a. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019b. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, Wei Shi, Yuliang Liu, Hao Liu, Yuan Xie, Xiang Bai, and Can Huang. 2024a. Textsquare: Scaling up text-centric visual instruction tuning. *CoRR*, abs/2404.12803.

Jingqun Tang, Chunhui Lin, Zhen Zhao, Shu Wei, Binghong Wu, Qi Liu, Hao Feng, Yang Li, Siqi Wang, Lei Liao, et al. 2024b. 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. 2024c. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model. *arXiv preprint arXiv:2202.13558*.

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*.