Lost in Translation and Noise: A Deep Dive into the Failure Modes of VLMs on Real-World Tables

Anshul Singh
Indian Institute of Science, Bangalore
anshulsinghchambial@gmail.com

Rohan Chaudhary
Panjab University, Chandigarh
chaudhary18.rohan@gmail.com

Gagneet Singh
Panjab University, Chandigarh
gagneet5647@gmail.com

Abhay Kumar
Panjab University, Chandigarh
abhaykumar.connect@gmail.com

Abstract

The impressive performance of VLMs is largely measured on benchmarks that fail to capture the complexities of real-world scenarios. Existing datasets for tabular QA, such as WikiTableQuestions and FinQA, are overwhelmingly monolingual (English) and present tables in a digitally perfect, clean format. This creates a significant gap between research and practice. To address this, we present **MirageTVQA**, a new benchmark designed to evaluate VLMs on these exact dimensions. Featuring nearly 60,000 QA pairs across 24 languages, MirageTVQA challenges models with tables that are not only multilingual but also visually imperfect, incorporating realistic noise to mimic scanned documents. Our evaluation of the leading VLMs reveals two primary failure points: a severe degradation in performance (over 35% drop for the best models) when faced with visual noise and a consistent English-first bias where reasoning abilities fail to transfer to other languages. MirageTVQA provides a benchmark for measuring and driving progress towards more robust VLM models for table reasoning. The dataset and the code are available at:https://github.com/anshulsc/MirageTVQA.

1 Introduction

Tables are the backbone of information storage in countless domains, from financial reports and scientific papers to enterprise databases and healthcare records. The ability to comprehend and reason over these semistructured data is an important skill for LLMs to act as an AI agent. Traditionally, answering questions from table images involved a cascading pipeline: first, an Optical Character Recognition (OCR) engine extracts text, which is then fed to a language model for reasoning. However, this multi-step process is highly susceptible to error propagation, where a single OCR mistake can invalidate the entire result. Recent work has shown that end-to-end vision language models (VLMs) can outperform these brittle OCR+LLM pipelines by jointly processing visual and textual information [Zheng et al., 2024].

This positions VLMs as a promising direction for robust table understanding. However, their true potential is hampered by a critical blind spot in existing evaluation benchmarks. Research has progressed along two largely separate tracks. On one hand, most benchmarks that explore complex reasoning, like FinQA [Chen et al., 2021], remain fundamentally **text-based**, ignoring the visual aspect of tables as they appear in documents. On the other hand, the few benchmarks that incorporate visual elements are almost exclusively **English-centric**. Crucially, to our knowledge, no existing benchmark addresses these two critical real-world challenges, visual complexity and linguistic diversity, **simultaneously**.

AI for Tabular Data workshop at EurIPS 2025

To fill this gap, we introduce MirageTVQA. Our primary contributions are twofold.

- We construct MirageTVQA, the first large-scale visual question-answering benchmark designed to test VLM reasoning on nearly 60,000 QA pairs. It integrates massive multilingual support (24 languages) with visually realistic table images.

- We conduct an extensive empirical evaluation of leading open-source VLMs, providing a comprehensive analysis of their performance across languages and their robustness against visual degradation.

2 Related Work

Our work builds upon and addresses the limitations of several lines of research in table understanding.

Text-Based Table Question Answering. A significant body of work has focused on reasoning over the textual content of tables. Benchmarks such as Spider [Yu et al., 2018], WikiTableQuestions [Pasupat and Liang, 2015], TAT-QA [Zhu et al., 2021], TableBench [Wu et al., 2025], and MIMO-Table [Li et al., 2025] frame the task as text-to-SQL or text-to-answer. These approaches typically feed models linearized HTML or Markdown renditions of tables, completely sidestepping the visual modality. While important for evaluating textual reasoning, they do not test a model's ability to process tables as they are found in documents.

Multi-modal Table Question Answering. More recent efforts have extended this research into the multi-modal domain. Work such as MMTab [Zheng et al., 2024] and MTabVQA[Singh et al., 2025] frame the task as visual table question answering, requiring models to reason over table images. However, these benchmarks have two key limitations: they remain monolingual (English-only), and the visual table images are synthetically generated and clean, lacking the real-world noise and artifacts often present in scanned documents or photographs.

Multilingual Table Question Answering. A third line of research has attempted to address the multilingual gap. Benchmarks like M3TQA [Shu et al., 2025] provide datasets for evaluation across multiple languages. Nevertheless, these efforts are often limited to a small number of language pairs (e.g., Chinese–English) and, critically, focus solely on text-based table representations, thereby inheriting the limitations of the first category. Our work is first to fill this gap by combining 20+ languages across diverse domains (scientific, financial, and general knowledge), evaluating a broader set of 10 reasoning types, introducing realistic visual noise to table images to create a benchmark to test VLM's tabular understanding.

3 MirageTVQA Benchmark

3.1 Data Collection and Table Translation

Source Table Collection and Filtering. We begin by collecting a diverse set of 3000 English tables from four primary sources: Wikipedia (WikiSQL [Zhong et al., 2017]), financial documents (FinQA [Chen et al., 2021]), scientific papers from arXiv, and GitHub. To ensure the suitability of tables for translation, we filtered the tables based on the median word character count per cell and selected a representative subset for translation. This process filtered our initial set down to 250 seed tables for the multilingual generation phase: 50 from arXiv, 100 from Wikipedia, and 100 from other sources, respectively. This curated set ensures a balanced and high-quality foundation for our benchmark.

With this curated seed, we create our multilingual Table corpus using a **translate-refine-filter** pipeline. For each of the 30 selected target languages, we: (1) perform an initial translation of all textual content using **Qwen3-32B** (Yang et al. [2025]); (2) use a powerful LLM (Gemini 2.5 Pro (Hassabis et al. [2024]) to refine the translation by cross-referencing the original English table for context and data integrity; (3) back-translate the refined table to English; and (4) filter out the languages that had low-quality translation on the basis of back-translation BLEU score (Papineni et al. [2002]), resulting in translated tables in 24 diverse languages.

3.2 Visually-Rich and Realistic Rendering

To address the reality gap, we developed a two-stage rendering pipeline. First, each table (in each language) is rendered into a clean PNG image from an HTML representation. We design 40+ distinct and advanced CSS themes to simulate a variety of document aesthetics, from minimalist and academic styles to financial reports and dark-mode interfaces. Next, we produce multiple noisy versions of each clean image. Using the `imgaug` library (Jung [2020]), we apply a stochastic pipeline of augmentations that are meant to simulate real-world degradation of documents. These involve geometric distortions, such as minor rotations, skew, and perspective transforms to simulate a camera capture; quality degradation, like Gaussian blur and variable JPEG compression; and the possibility of scanning artifacts, such as salt-and-pepper noise, slight scan lines, and corner shadowing. This process results in a rich dataset where each table is associated with one clean image and multiple unique noisy counterparts, complete with metadata for each applied transformation.

3.3 Question-Answer Generation

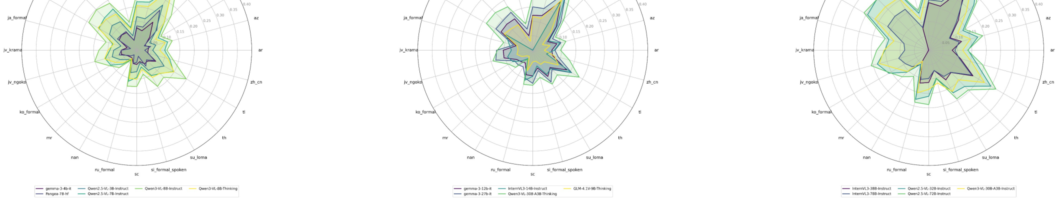
We produced question-answer pairs using a combined human-LLM approach to balance cost efficiency and quality of the annotation process. First, human annotators curated a seed corpus by manually creating one high-quality QA pair per table. Next, we expanded the QA dataset by utilizing the seed examples in detailed prompts, allowing LLMs (Google Gemini (Hassabis et al. [2024]) to produce 10 additional QA pairs for each table, covering 10 reasoning types and two question types. This process yielded 11 QA pairs for each table-language combination, for a total of 80,520 QA pairs (244 tables × 30 languages × 11 QA pairs). Following the generation of the QA pairs, we validated these pairs with LLMs and identified which pairs had been misclassified during generation. Three human annotators then corrected these flagged pairs. With this final correction, the QA pairs for each of the tables were complete. Both source tables and their associated QA pairs were translated into all target languages to enable extensive multilingual evaluation. For more details, refer to Appendix A.

4 Results and Analysis

Our benchmark, MirageTVQA, was designed to evaluate Vision-Language Models (VLMs) on two critical axes that reflect real-world challenges: **visual robustness** and **multilingual reasoning**. To do this, our dataset provides two visual settings for each table: digitally perfect clean images and noisy images that simulate document degradation. The analysis of model performance on MirageTVQA reveals significant and systematic limitations in current state-of-the-art models.

4.1 Baseline Performance and Model Scale

We first establish a performance baseline by evaluating models on the clean image set. The results, detailed in Table 1, demonstrate a strong and consistent correlation between model scale and reasoning capability. The largest model evaluated, Qwen2.5-VL-72B (Team et al. [2025c]), achieves the highest average Exact Match (EM) of 13.57% across all languages, significantly outperforming smaller models. This trend, visually represented by the expanding area of the performance polygons in Figure



(a) Small (3B-8B)

(b) Medium (9B-30B)

(c) Large (32B-78B)

Figure 1: Per-language F1 score performance on **MirageTVQA** across different model scales.

Table 1: Comprehensive per-language performance breakdown on **MirageTVQA**, showing Exact Match (%). The 'Avg.' column shows the overall average EM. I means Instruct models and T means Thinking models. Best result in each column is in **bold**.

Size	Model	T	Avg	en	es	fr	it	cs	ru	tl	id_f	id_c	su	jv_k	jv_n	az	th	bn	si	ja	ar	sc	zh	hi	ko	mr	nan
<10B	Pangea-7B (Yue et al. [2024])	I	1.03	5.23	1.29	1.29	1.36	3.62	1.00	0.92	1.31	1.27	0.80	0.62	0.83	0.71	0.66	0.41	0.21	0.25	0.36	0.37	0.62	0.49	0.41	0.21	0.12
	Qwen2.5-3B (Team [2024])	I	1.83	5.19	2.75	2.90	2.38	2.10	2.37	2.05	2.88	2.63	1.39	1.44	1.53	1.13	2.06	1.15	0.82	1.44	0.81	1.10	1.32	1.51	1.60	1.11	0.12
	Gemma-3-4B (Team et al. [2025a])	I	1.78	3.67	2.34	2.45	2.34	1.77	1.91	2.05	2.51	2.14	1.47	1.52	1.53	1.46	1.60	1.60	1.52	1.52	1.45	1.43	1.44	1.47	1.31	1.39	0.80
	Qwen2.5-7B (Team [2024])	I	4.57	9.85	6.93	6.43	5.59	5.15	5.77	4.81	6.04	6.82	3.29	2.30	2.98	4.26	5.14	3.69	1.60	4.28	4.71	2.86	5.14	3.79	4.79	2.34	0.88
	Qwen3-8B (Yang et al. [2025])	I	8.01	17.53	10.93	10.00	10.72	9.68	7.63	10.33	10.44	10.06	7.37	6.58	7.48	6.52	8.19	5.86	4.64	7.74	6.20	6.55	8.19	6.40	5.98	5.00	1.73
	GLM-4.1V-9B (Hong et al. [2025])	T	2.73	8.17	4.63	4.11	4.85	3.01	2.95	4.39	4.23	4.15	1.85	1.73	2.07	2.17	1.89	1.43	1.44	1.94	1.61	1.72	2.59	1.39	1.80	0.78	0.44
	Qwen3-8B (Yang et al. [2025])	T	6.36	15.55	8.97	8.50	8.92	7.66	6.31	7.65	8.05	7.89	4.80	4.61	5.58	5.52	6.58	4.88	3.41	5.48	4.83	5.11	7.37	4.69	5.04	3.61	1.17
Med	Gemma-3-12B (Team et al. [2025a])	I	5.31	10.66	6.97	6.43	6.86	6.67	5.73	6.86	6.94	6.57	5.35	5.10	4.88	4.97	4.86	4.02	3.78	4.24	4.35	4.58	4.20	3.18	4.01	3.57	2.41
	InternVL3-14B (Zhu et al. [2025])	I	3.69	12.72	0.00	0.00	5.26	6.34	6.31	7.19	0.00	0.00	3.92	2.63	3.72	4.68	3.91	3.40	2.71	5.68	4.23	4.09	3.37	0.00	4.06	2.54	1.13
	Gemma-3-27B (Team et al. [2025a])	I	6.71	13.79	9.06	8.59	8.83	8.11	6.51	8.70	8.34	8.67	6.62	6.66	6.66	5.77	5.10	5.33	4.07	4.74	5.16	6.63	5.72	5.46	4.59	4.59	3.06
	Qwen3-30B (Yang et al. [2025])	I	9.45	20.05	12.27	12.86	12.16	10.71	9.71	11.54	11.59	11.50	8.51	7.77	9.38	8.40	8.64	7.21	6.86	9.27	7.21	6.55	10.66	7.01	8.19	6.32	2.05
	Qwen3-30B (Yang et al. [2025])	T	8.19	18.95	11.44	11.07	11.18	9.18	7.76	9.82	10.48	10.14	6.87	6.33	7.85	6.90	7.33	5.82	5.92	7.37	5.92	6.91	9.26	6.40	6.02	5.33	1.85
Large	InternVL3-38B (Zhu et al. [2025])	I	4.76	15.16	9.31	9.25	9.16	8.48	7.55	10.20	8.50	8.34	6.15	-	-	7.02	6.01	5.49	3.90	-	4.67	5.97	7.61	4.93	-	-	-
	InternVL3-78B (Zhu et al. [2025])	I	11.22	27.84	17.53	17.47	16.47	13.50	10.48	17.25	16.65	16.40	10.51	8.30	10.80	10.98	10.83	8.31	7.26	8.9	9.16	9.80	11.02	8.16	9.35	8.08	5.80
	Qwen2.5-72B (Team [2024])	I	**13.57**	**25.52**	**17.57**	**17.42**	**18.08**	**15.94**	**14.07**	**17.52**	**17.01**	**16.84**	**12.22**	**10.98**	**12.86**	**13.21**	**13.50**	**10.66**	**6.54**	**11.90**	**10.07**	**12.15**	**14.28**	**10.80**	**11.27**	**9.39**	**5.47**

1, confirms that greater parameter counts are beneficial for the complex, multi-step reasoning required by MirageTVQA.

4.2 The Impact of Visual Noise

A primary motivation for creating MirageTVQA was to measure how models performs with the visual imperfections common in real-world documents. Our findings reveal that current models are extremely brittle in this regard. Table 2 directly compares model performance on 'clean' versus 'noisy' images for the English subset. The top-performing Qwen2.5-VL-72B (Team et al. [2025c]) model, which scores 25.52% EM on clean images, sees its performance degrade to just 16.50% EM on the noisy set with drop of over **35%**. This trend is consistent across all evaluated models. Hence, it validates a core premise of our benchmark: that performance on pristine, synthetic data is an unreliable predictor of performance on realistic, visually imperfect data.

4.3 Multilingual Performance

The second core motivation of MirageTVQA was to assess reasoning capabilities beyond English. Our results show the VLMs' performance is biased towards English (see Figure 1). Across all model scales, performance invariably peaks on English. Scores drop sharply even for other high-resource languages, degrade further for languages with different scripts, and become negligible for many low-resource languages. This demonstrates that current VLMs, despite their exposure to multilingual data during pre-training, fail to generalize complex, visually grounded reasoning skills to non-English contexts. This failure in cross-lingual transfer is a critical gap that MirageTVQA successfully exposes, reinforcing the need for more equitable and truly multilingual model development.

Table 2: Impact of Visual Noise on Model Performance (EM %) on the English Subset. Models are sorted by their performance on clean images.

Model	Clean EM (%)	Noisy EM (%)	Perf. Drop (%)
Gemma-3 27B-IT (Team et al. [2025b])	13.79	12.87	-6.7%
Qwen-2.5-VL 8B-Instruct (Team et al. [2025c])	17.53	16.62	-5.2%
Qwen-2.5-VL 32B-Instruct (Team et al. [2025c])	23.15	20.36	-12.1%
Qwen-2.5-VL 72B-Instruct (Team et al. [2025c])	**25.52**	**16.50**	**-35.3%**

5 Conclusion

In this work, we introduced MirageTVQA, a new large-scale benchmark designed to address a critical gap in the evaluation of vision-language models: their ability to reason over tables that are both visually imperfect and multilingual. Through our extensive experiments, we have demonstrated two significant findings. First, the performance of even state-of-the-art VLMs degrades in the presence of realistic visual noise, highlighting a profound lack of robustness. Second, we identified a severe "English-first" bias, with models failing to transfer their reasoning capabilities to non-English and low-resource languages. We hope it will be useful for the development of more robust and capable models for real-world table understanding. Additionally, we discuss the limitations of our study and outline future directions toward improving the interpretability of VLM failure modes in the Appendix.

References

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Demis Hassabis, Koray Kavukcuoglu, and Google DeepMind. Introducing Gemini 2.0: Our New AI Model for the Agentic Era. Blog post, Google DeepMind, December 11 2024. URL `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025.

Alexander Jung. imgaug - image augmentation for machine learning. `https://github.com/aleju/imgaug`, 2020.

Zheng Li, Yang Du, Mao Zheng, and Mingyang Song. MiMoTable: A Multi-scale Spreadsheet Benchmark with Meta Operations for Table Reasoning. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2548–2560, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.coling-main.173/`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040/`.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.

Daixin Shu, Jian Yang, Zhenhe Wu, Xianjie Wu, Xianfu Cheng, Xiangyuan Guan, Yanghai Wang, Pengfei Wu, Tingyang Yang, Hualei Zhu, et al. M3TQA: Massively multilingual multitask table question answering. *arXiv preprint arXiv:2508.16265*, 2025.

Anshul Singh, Chris Biemann, and Jan Strich. MTabVQA: Evaluating multi-tabular reasoning of language models in visual space. *arXiv preprint arXiv:2506.11684*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025a.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel

Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 Technical Report, 2025b. URL https://arxiv.org/abs/2503.19786.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2025c. URL https://arxiv.org/abs/2412.15115.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Tongliang Li, Zhoujun Li, and Guanglin Niu. TableBench: A Comprehensive and Complex Benchmark for Table Question Answering. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25, pages 25497–25506. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i24.34739. URL https://doi.org/10.1609/aaai.v39i24.34739.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*, 2018.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2024.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal Table Understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages

9102–9124, Bangkok, Thailand3, 2024. Association for Computational Linguistics. doi: 10.18653/V1/2024.ACL-LONG.493. URL https://doi.org/10.18653/v1/2024.acl-long.493.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning, 2017. URL https://arxiv.org/abs/1709.00103.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Technical Appendices and Supplementary Material

A.1 Data distribution and Language Composition

To provide a transparent overview of our benchmark, Table 3 details the composition of MirageTVQA. Part (a) shows the initial breakdown of QA pairs and unique tables collected from our diverse sources before filtering and translation. Part (b) shows the final distribution of the nearly 60,000 high-quality QA pairs that passed our validation pipeline, broken down by language and grouped by linguistic family. This highlights the broad and balanced coverage of our final benchmark.

Table 3: Detailed composition of the **MirageTVQA** benchmark. The table shows the distribution of tables and QA pairs by their original source, followed by the final distribution of QA pairs per language, grouped by linguistic family.

(a) Data Source Composition

Data Source	# QA Pairs	Unique Tables	Total Table Images
ArXiv	12,100	1,100	4,400
Wikipedia	27,500	2,500	10,000
Other (Financial, etc.)	27,500	2,500	10,000
Total Initial Set	**67,100**	**6,100**	**24,400**

(b) Language Composition (Final Valid Set)

Language Family	Language	# Final QA Pairs
Afro-Asiatic	Arabic (ar)	2,482
Austronesian	Indonesian (id_casual)	2,435
	Indonesian (id_formal)	2,434
	Javanese (jv_krama)	2,431
	Javanese (jv_ngoko)	2,419
	Sundanese (su_loma)	2,373
	Tagalog (tl)	2,392
Indo-European	Bengali (bn)	2,440
	Czech (cs)	2,428
	English (en)	2,618
	French (fr)	2,411
	Hindi (hi)	2,454
	Italian (it)	2,434
	Marathi (mr)	2,438
	Russian (ru_formal)	2,410
	Sardinian (sc)	2,444
	Sinhala (si_formal_spoken)	2,433
	Spanish (es)	2,396
Japonic	Japanese (ja_formal)	2,428
Koreanic	Korean (ko_formal)	2,441
Kra-Dai	Thai (th)	2,430
Sino-Tibetan	Hokkien (nan)	2,487
	Chinese (zh_cn)	2,430
Turkic	Azerbaijani (az)	2,392
Total Valid Set	**24 Languages**	**58,480**

A.2 Table Translation Prompt

For reproducibility, we provide the exact prompts used in our data generation pipeline. Figure 3 shows the detailed prompt provided to Gemini 1.5 Pro (Comanici et al. [2025])to generate the English dense-reasoning question-answer pairs described in Section 3.3. Figure 2 shows the prompt used to translate these QA pairs into the 25 target languages.

The following prompt shown in Fig 2 was provided to the LLMs QA translation agents during the QA translation phase described in Section 3.1.

```
You are an expert linguist and professional translator with deep expertise
in structured data.  Your task is to accurately translate a question-answer
pair from English to {target_language}.
The question-answer pair is based on the following data table.  Use this
table to understand the context of entities, numbers, and technical terms.

Context Table:
{context_table_json}
----------------
ENGLISH QUESTION-ANSWER PAIR TO TRANSLATE:
{english_qa_json}
----------------
CRITICAL INSTRUCTIONS:
1.  Convert the question text into fluent and natural-sounding
{target_language}.
2.  If the question_type is 'open_ended_reasoning', you must translate the
full text of the answer.  However, if the question_type is 'value', you
must preserve the original answer exactly.  Do not translate numbers (e.g.,
"370"), names (e.g., "Beta"), percentages, or codes.
3.  Your entire response must be a single, valid JSON object with ONLY two
keys:  translated_question and translated_answer.  Do not add any other
text, explanations, or markdown code blocks.
---------------
EXAMPLE:
If target_language is Spanish and the input QA pair is:
{
  "question":  "Which product experienced a decline in units sold from 2022
to 2023?",
  "answer":  [["Beta"]],
  "question_type":  "value"
}

Your perfect JSON output would be:
{
  "translated_question":  "¿Qué producto experimentó una disminución en las
unidades vendidas de 2022 a 2023?",
  "translated_answer":  [["Beta"]]
}

Notice how "Beta" was NOT translated because the question_type was 'value'.
--

YOUR TASK:
Translate the provided English QA pair to {target_language} following all
instructions.

Your JSON Output:
```

Figure 2: LLM prompt for multilingual QA pair translation. Placeholders like `{target_language}` and `{context_table_json}` represent actual input data provided to the model.

A.3 QA pair generation prompt

The following prompt shown in Fig 3 was provided to the LLMs QA generation agents during the QA generation phase described in Section 3.3.

```
You are a world-class data analyst and expert curriculum designer.
Your task is to generate a set of {num_questions} diverse, challenging, and
high-quality question-answer pairs based on the provided data table in JSON
format.
The questions must require deep reasoning and not be simple lookups.


CRITICAL INSTRUCTIONS:
1.  Create questions that cover a wide range of the reasoning categories
defined below.  Do not repeat question patterns.
2.  Every question must be answerable exclusively from the provided table.
Do not require external knowledge.
3.  Provide Precise Answers:
  - For value-based questions:  The ’answer’ must be the exact value(s)
from the table or calculated from it.  Format it as a list of lists (e.g.,
[["150"]] or [["Alpha"], ["Beta"]]).
  - For open-ended reasoning questions:  The ’answer’ should be a
comprehensive explanation or analysis based on the data, formatted as a list
containing a single list with one string element (e.g., [["The data shows a
declining trend because..."]]).
4.  Each question must have a ’question_type’ field that is either "value" or
"open_ended_reasoning".
5.  The ’evidence_cells’ must accurately list all cells needed to
formulate the answer.  Use standard spreadsheet notation (e.g., A1 for the
top-left-most cell in the data body, where Column A is the first column and
Row 1 is the first data row.  Headers are considered Row 0, so A1 refers to
the first data cell, not a header).
6.  Your response MUST be a single, valid JSON object that conforms to the
provided schema.  Do not include any explanatory text, markdown, or comments
outside of the JSON structure, and do not wrap the JSON in markdown code
blocks (e.g., ``ǰson).


REASONING CATEGORIES TO USE: Comparative Reasoning, Numerical Aggregation,
Multi-Hop Reasoning, Temporal Reasoning, Conditional Reasoning,
Proportional/Ratio Analysis, Hypothetical Reasoning, Correlation Inference,
Structural/Metadata Reasoning, Outlier Detection

REQUIRED JSON OUTPUT SCHEMA:
{
  "qa_pairs":  [
    {
      "question":  "string",
      "answer":  [["string"]],
      "evidence_cells":  ["string"],
      "reasoning_category":  "string (must be one of the 10 categories)",
      "question_type":  "string (either ’value’ or ’open_ended_reasoning’)"
    }
  ]
}


NOW, GENERATE QA PAIRS FOR THE FOLLOWING TABLE:
Input Table (as JSON): {table_as_json_string}
Your JSON Output:
```

Figure 3: LLM prompt for automated QA pair generation. Placeholders like {table_as_json_string} represent the actual table data provided to the model.

B Limitations

While this research provides an understanding of how real-world noise affects the performance of VLMs in multi-modal visual question-answers, we would like to emphasize a few limitations. First, we did analyze how noise impacts performance, but we did not establish interpretability methods to explain why these degradations occur or suggest ways to reduce that degradation. This will be an important area for future work. Second, we carried out our cross-lingual experiments in a limited scope. The cross-lingual experiments involved performance in only 25 languages, we only evaluated open models, and we did not perform experiments on top-tier proprietary models that may behave differently on robustness in noise. Adopting state-of-the-art proprietary models, introducing a wider array of languages, and exploring interpretability methods for understanding and addressing degradation are all potential improvements to our analyses.