Systematic Performance Degradation in Indic Vision-Language Models: Evidence from Hindi and Telugu

Anonymous ACL submission

Abstract

With 1.5 billion people speaking over 120 major languages, India exemplifies the challenges of multilingual AI evaluation. Current multilingual VLM benchmarks suffer from unverified auto-translations, narrow task coverage, small sample sizes, and lack of culturally grounded content. We present HinTel-AlignBench, a comprehensive evaluation framework and benchmark for Hindi and Telugu vision-language models with English-aligned samples. Our framework combines semi-automated translation with human verification to generate ∼4k QA pairs per language across five domains: adapted English datasets (VQAv2, RealWorldQA, CLEVR-Math) and native Indic sets (JEE for STEM, VAANI for cultural grounding). Evaluation of state-of-the-art open and closed-source VLMs reveals consistent performance regression from English to Indic languages, with average drops of 8.3 points for Hindi and 5.5 points for Telugu across four of five tasks. We identify key failure modes and establish reproducible baselines for multilingual multimodal evaluation.

1 Introduction

India's 122 major languages and 1599 other languages[1] present unique challenges for multilingual AI. While recent multimodal large language models (MLLMs) such as ChatGPT (OpenAI, 2025), Gemini 2.5 (Google DeepMind, 2025), and open-weight variants (Meta Llama, 2025; Dash et al., 2025) claim multilingual support, comprehensive evaluation benchmarks for Indian languages remain scarce.

Current evaluation methodologies suffer from critical limitations in quality, scope, and scale. First, many benchmarks rely on unverified automatic translations (Wu et al., 2025), inevitably introducing noise. While text-only benchmarks like IndicGenBench (Singh et al., 2024) exist, they lack multimodal coverage. Second, existing vision-language benchmarks often suffer from insufficient sample sizes; for instance, xChat (Yue et al., 2025) and AyaVisionBench (Dash et al., 2025) contain only 50 and 135 QA pairs per language, respectively, preventing statistically significant analysis. Third, domain coverage is often narrow. Concurrent work such as Kaleidoscope (Salazar et al., 2025) provides ∼800 samples per Indic language but focuses exclusively on exam-based multiple-choice questions, neglecting real-world reasoning. Finally, adapted benchmarks often lack cultural grounding, evaluating surface-level translation rather than native competence (Khan et al., 2024). We elaborate on previous works in the appendix.

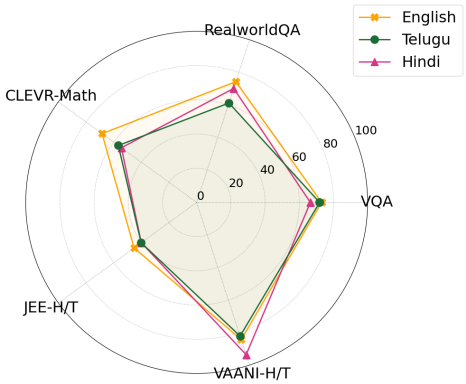

Figure 1: Average performance of GPT-4.1 and Gemini-2.5-Flash on English, Hindi, and Telugu across data-parallel visual question answering samples. Performance regresses from English to Hindi by 8.3 points and from English to Telugu by 5.5 points.

To address these gaps, we introduce HinTel-AlignBench, a scalable framework and benchmark for evaluating VLMs in Hindi and Telugu. Our semi-automated pipeline combines translation or LLM-based QA generation with strict human verification, achieving 5x faster processing than manual

[1] https://en.wikipedia.org/wiki/Languages_of_India

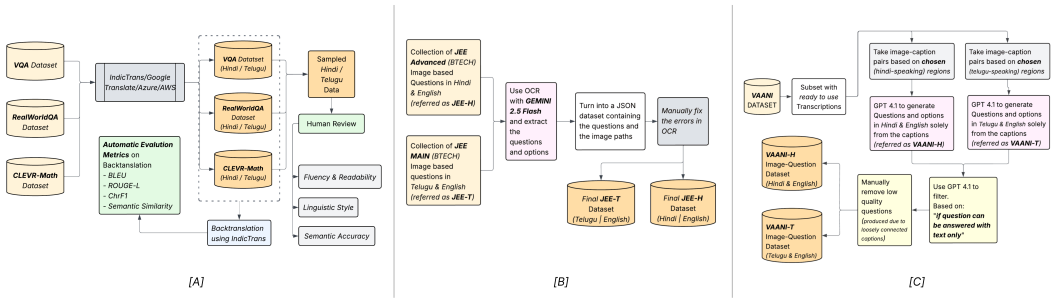

Figure 2: Dataset generation pipeline for (A) VQAv2, RealWorldQA, and CLEVR-Math using translation and human verification; (B) JEE-H and JEE-T using OCR extraction and verification; (C) VAANI-H and VAANI-T using LLM-based question generation from captions with filtering and verification.

creation for 79% of samples while maintaining linguistic fidelity. The benchmark comprises ∼4k QA pairs per language—significantly larger than prior manually verified sets—spanning five domains: real-world understanding (VQAv2 (Goyal et al., 2017a)), practical reasoning (RealWorldQA (xAI, 2024a)), visual mathematics (CLEVR-Math (Lindström and Abraham, 2022)), STEM competency (JEE-Vision from India's Joint Entrance Exam), and cultural grounding (VAANI (Team, 2025)). Crucially, each sample includes manually verified English translations, enabling direct cross-lingual comparison.

Evaluation of state-of-the-art models on our benchmark reveals systematic performance degradation. Across all models, we observe average regressions of 8.3 points (Hindi) and 5.5 points (Telugu) relative to English, with gaps appearing in four of five tasks (Figure 1). Even frontier models like GPT-4.1 exhibit 3.8-point (Hindi) and 8.6-point (Telugu) performance drops. Performance on aligned Hindi and Telugu subsets differs by less than 1 point, indicating comparable gaps between English and both Indic languages.

Our contributions are: (1) a semi-automated framework for generating multilingual vision-language evaluation sets; (2) the largest human-verified Hindi and Telugu VLM benchmark to date, featuring culturally sourced content and English-aligned samples; and (3) a comprehensive evaluation of state-of-the-art models, highlighting significant performance regressions across diverse domains.

2 Datasets

2.1 Data Sources

We construct HinTel-AlignBench by combining translated English VQA datasets with native Indic evaluation sets across five domains. The translated sets include 1000 samples from VQAv2 (Goyal et al., 2017b) for real-world visual understanding, 765 samples from RealWorldQA (xAI, 2024a,b) for practical spatial reasoning, and 1000 samples from CLEVR-Math (Lindström and Abraham, 2022) for visual mathematical reasoning.

For native Indic content, we develop JEE-Vision from India's Joint Entrance Examination, sourcing 192 Hindi questions (JEE-H) from the Advanced exam and 325 Telugu questions (JEE-T) from the Mains exam. These diagram-dependent STEM problems span mathematics, physics, and chemistry, providing the first benchmark for non-translated multilingual technical reasoning with visual content. We generate culturally grounded evaluation sets (VAANI-H and VAANI-T) by sampling 945 Hindi and 1020 Telugu images from the VAANI corpus (Team, 2025), using GPT-4.1 to create multiple-choice questions from original captions, then filtering out text-only answerable questions.

Translation-based extension enables multi-way parallel data, allowing attribution of performance to task knowledge versus language understanding (Singh et al., 2024). This approach also leverages the quality control invested in designing the original English benchmarks. Table 1 shows the distribution of QA pairs per language and task. Figure 3 showcases a few examples.

Language	VQAv2	RealWorldQA	CLEVR-Math	JEE-H	JEE-T	VAANI-H	VAANI-T
Hindi	1,000	765	1,000	192	-	945	-
Telugu	1,000	765	1,000	-	325	-	1,020
English	1,000	765	1,000	192	325	945	1,020
Total	3,000	2,295	3,000	384	650	1,890	2,040

Table 1: Number of QA pairs per task per language in HinTel-AlignBench. The samples used in VQAv2, RealWorldQAand CLEVR-Math are the same across all languages.

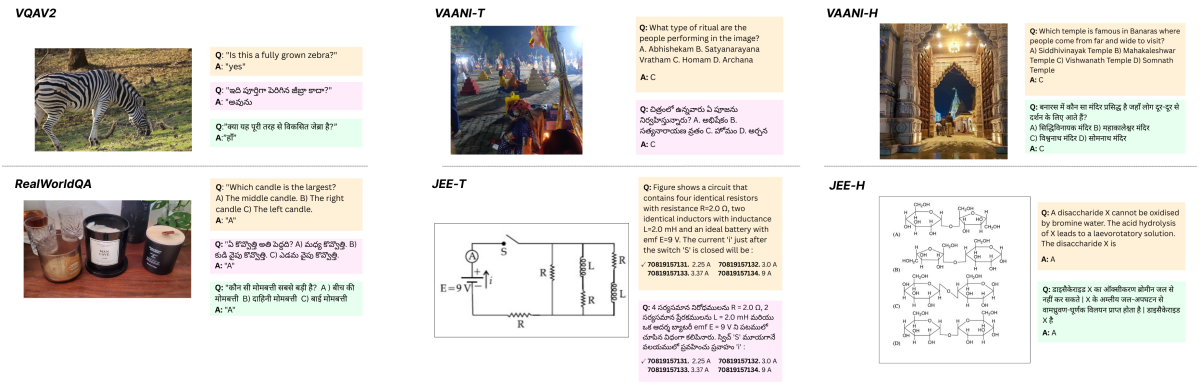

Figure 3: Qualitative Examples for different domains in our dataset. More images are shown in the appendix

2.2 Dataset Generation Framework

Figure 2 illustrates our three-stage generation framework tailored to different data sources.

Translation Pipeline. For VQAv2, RealWorldQA, and CLEVR-Math, we evaluate four translation systems (IndicTrans (Gala et al., 2023), Google Translate, Azure, AWS) on 50 diverse samples per language, selecting Azure for Hindi and AWS for Telugu. All translations undergo manual verification for semantic accuracy, linguistic style, and readability (KJ et al., 2025). We avoid back-translation-based sample selection, which introduces bias toward high-confidence translation errors. Manual review accepts 79% of VQAv2 translations without modification; among modified samples, 42% require only minor changes (verb tense), while 58% need word addition or deletion. Samples requiring only minor edits process 5x faster than generation from scratch. All verification is performed by co-author native speakers. We use one annotator per sample to maximize dataset size within budget constraints.

JEE-Vision Creation. India's Joint Entrance Examination provides authentic STEM problems authored by subject-matter experts in target languages (JEE Mains: 13 languages; Advanced: English/Hindi), avoiding translation artifacts. We curate diagram-dependent problems, evaluating joint understanding of technical visuals and linguistic content. Questions and options are extracted using Gemini-2.5-Flash OCR (Google DeepMind, 2025), then manually verified to correct OCR errors.

VAANI Generation. From the VAANI corpus (Team, 2025), we extract images with text transcriptions from Hindi and Telugu speaking regions. Since no images have both Hindi and Telugu transcriptions, we create separate language-specific sets. Text-only GPT-4.1 generates multiple-choice questions from captions, which undergo two-stage refinement: automated filtering removes questions answerable without images, followed by human verification to eliminate low-quality questions. This process addresses cases where VAANI captions do not perfectly align with images.

3 Experimental Setup

Models. We evaluate open-weight and proprietary models claiming Indic language support. For Hindi, we test Gemma3 (4B, 12B, 27B) (Team et al., 2025), Qwen2.5VL-7B (Bai et al., 2025), Llama3.2-Vision-11B (Meta Llama, 2025), Aya-8B (Dash et al., 2025), Chitrarth-8B (Khan et al., 2024), GPT-4.1 (OpenAI, 2025), and Gemini-2.5-Flash (Google DeepMind, 2025). For Telugu, fewer models provide support; we evaluate the Gemini variants, GPT-4.1, and Chitrarth-8B.

Metrics. For multiple-choice tasks (RealWorldQA, VAANI, JEE), we report standard accuracy. For open-ended generation (VQAv2, CLEVR-Math), we utilize a hybrid evaluation protocol combining exact match with a GPT-4.1 judge to account for linguistic variations, following standard VQA practices (complete details in appendix).

4 Results and Analysis

4.1 Main Results

Tables 2 and 3 present Telugu-English and Hindi-English comparisons. Across all models and tasks, average performance regresses 5.5 points from English to Telugu and 8.3 points from English to Hindi. Performance drops occur in four of five tasks, with VAANI showing smaller gaps.

Model	VQAv2		RealWorldQA		CLEVR-Math		JEE-T		VAANI-T		Ours-T	
	Tel	En	Tel	En	Tel	En	Tel	En	Tel	En	Tel	En
GPT 4.1	68.70	72.00	61.05	**75.29**	46.60	65.00	34.36	40.92	81.57	**82.45**	58.46	67.13
Gemini 2.5 Flash	75.10	74.10	**61.18**	73.20	**66.70**	**71.80**	**45.90**	**54.15**	**82.75**	80.78	**66.33**	**70.81**
Gem 2.0 Flash	70.20	74.20	60.92	69.67	43.60	53.50	42.15	53.23	80.49	79.61	59.47	66.04
Gem 1.5 Flash	68.50	74.40	60.00	67.19	37.40	46.70	29.85	39.38	76.27	79.61	54.40	61.46
Chitrarth	**76.00**	**78.50**	53.59	52.55	53.90	56.90	20.00	18.15	81.57	82.45	57.01	57.71
Model Mean	71.10	74.64	59.35	67.58	49.64	58.78	34.85	41.17	80.53	80.98	59.13	64.63

Table 2: Results (in %) for Telugu and English. **Bold** indicates the best and underline indicates the next best.

Model	VQAv2		RealWorldQA		CLEVR-Math		JEE-H		VAANI-H		Ours-H	
	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En
GPT-4.1	**68.00**	72.00	**70.59**	**75.29**	48.10	65.00	23.18	23.05	93.33	86.88	60.64	64.44
Gemini 2.5 Flash	65.00	74.10	69.54	73.20	**60.70**	71.80	**56.90**	**62.89**	**93.86**	**87.19**	**69.20**	**73.84**
Chitrarth	66.00	**78.50**	52.94	52.55	57.20	56.90	11.72	13.93	84.23	80.14	54.42	56.40
Qwen2.5VL-7B	37.20	74.30	51.11	68.10	29.10	**98.80**	17.84	20.70	81.79	84.76	43.41	69.33
Aya-8B	36.30	47.30	55.42	58.82	46.20	61.40	9.63	16.02	82.22	80.42	45.95	52.79
LLaMA 3.2 11B	35.90	59.80	35.68	61.57	18.90	35.60	14.32	14.45	77.67	83.28	36.49	50.94
Gemma3-27B	64.10	65.50	54.38	61.04	43.50	53.70	19.66	17.58	87.41	82.01	53.81	55.97
Gemma3-12B	63.00	65.70	53.98	58.69	40.20	46.80	14.84	16.80	85.50	82.33	51.50	54.87
Gemma3-4B	55.00	58.20	43.27	50.19	33.20	39.60	14.32	17.19	80.64	77.78	45.29	48.59
Model Mean	53.74	66.26	53.99	62.49	43.01	58.62	20.01	22.29	85.85	83.76	51.19	58.49

Table 3: Results (in %) for Hindi and English. **Bold** indicates the best and underline indicates the next best.

On aligned samples spanning VQAv2, CLEVR-Math, and RealWorldQA evaluated with GPT-4.1, Gemini-2.5-Flash, and Chitrarth, Hindi, Telugu, and English achieve 61.51, 62.53, and 68.6 points respectively, demonstrating systematic degradation from English to both Indic languages.

Gemini-2.5-Flash achieves best overall performance on both language pairs. Chitrarth leads on VQAv2 due to multilingual VQAv2 training. However, models show substantial cross-language variance: GPT-4.1 excels on RealWorldQA in English and Hindi but underperforms on Telugu, highlighting the need for comprehensive evaluation across all target languages. Qwen2.5VL-7B exhibits the largest Hindi-English gap at 25.92 points.

4.2 Task-Specific Analysis

Figure 4 shows average performance regression per task. CLEVR-Math and RealWorldQA exhibit the largest English-Indic gaps, while VAANI shows the smallest. VAANI-H performance exceeds English by 2.09 points on average. Analysis reveals two factors: first, some English questions fail to capture Indic-script option meanings in images with visible Indic text. Second, text-only LLM-generated distractors may enable statistical pattern exploitation.

Chain-of-Thought prompting improves reasoning tasks but benefits English more than Hindi. On JEE-H, CoT gains 13.8 points in English versus 2.22 points in Hindi, suggesting training bias toward English CoT data. Detailed ablationions and error analysis are reported in the appendix.

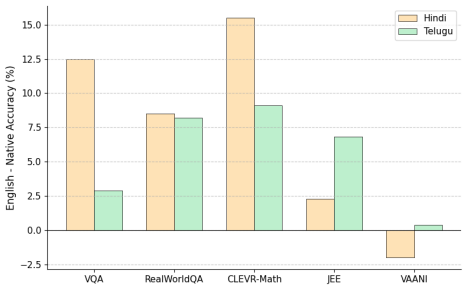

Figure 4: Average performance regression from English to Indic languages per domain. All tasks except VAANI show consistent regression.

5 Conclusion

This paper introduces HinTel-AlignBench, a framework for developing benchmarks to evaluate multimodal large language models in Hindi and Telugu, addressing critical gaps in existing multilingual evaluations. We combined semi-automated dataset creation with rigorous human verification and sourced culturally grounded native datasets to assess diverse capabilities. Evaluations of state-of-the-art VLMs reveal significant performance regressions in Indic languages compared to English, emphasizing the need for targeted improvements in multilingual visual understanding.

Limitations

While our benchmark introduces a diverse evaluation set it has limitations. First, the proprietary models we evaluated achieve high scores on the VAANI-H/T. We use text only LLMs for generating QA from VAANI captions and they often do not design good distractors. Thus, the models may guess the correct answer by exploiting statistical patterns, which inflates metrics. A future work is using Multi-Binary Accuracy (Cai et al., 2024) for VAANI subsets. Second, the JEE-H benchmark contains only 2 Mathematics questions, due to lack of image-based mathematics questions in the JEE-Advanced examination. Finally, there are 22 official Indian languages and we cover English and 2/22 (Hindi and Telugu) with this work. We hope this benchmark gets extended to all other Indic languages with contributions from native speakers from those languages.

References

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, and 1 others. 2025. Behind maya: Building a multilingual vision language model. *arXiv preprint arXiv:2505.08910*.

Anthropic. 2025. Claude 3.7 sonnet: The first hybrid reasoning model. Anthropic Company Website. Available at: https://www.anthropic.com/news/claude-3-7-sonnet.

Daman Arora, Himanshu Singh, and Mausam. 2023. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. Temporalbench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

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 *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*, arXiv:2505.08751.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent ai model. Google DeepMind Blog. Released on March 25, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. 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 (CVPR)*, pages 6904–6913.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Confer-*

ence on Computer Vision and Pattern Recognition (CVPR).

Shaharukh Khan, Ayush Tarun, Abhinav Ravi, Ali Faraz, Akshat Patidar, Praveen Kumar Pokala, Anagha Bhangare, Raja Kolla, Chandra Khatri, and Shubham Agarwal. 2024. Chitrarth: Bridging vision and language for a billion people. In *NeurIPS Multimodal Algorithmic Reasoning*.

Sankalp KJ, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. 2025. Indicmmlu-pro: Benchmarking indic large language models on multi-task language understanding. *arXiv preprint arXiv:2501.15747*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual, and mathematical reasoning. *arXiv preprint*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Meta Llama. 2025. Llama 3.2-vision: Instruction-tuned image reasoning generative models. Model release by Meta. Available at: https://huggingface.co/meta-llama/Llama-3.2-11B-Vision.

OpenAI. 2025. Gpt-4.1: A new series of gpt models with major improvements on coding, instruction following, and long context. OpenAI Company Website. Released on April 14, 2025.

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

Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2025. Palo: A polyglot large multimodal model for 5b people. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 1745–1754.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 56 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, Dominik Krzemiński, Jekaterina Novikova, Luísa Shimabucoro, Joseph Marvin Imperial, Rishabh Maheshwary, Sharad Duwal, Alfonso Amayuelas, Swati Rajwal, Jebish Purbey, and 25 others. 2025. Kaleidoscope: In-language exams for massively multilingual vision evaluation. *Preprint*, arXiv:2504.07072.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, and 1 others. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

VAANI Team. 2025. Vaani: Capturing the language landscape for an inclusive digital india (phase 1). https://vaani.iisc.ac.in/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*, arXiv:2504.15521.

xAI. 2024a. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v.

xAI. 2024b. Realworldqa dataset. Hugging Face Dataset Repository.

XAI.org. 2024. Realworldqa: A benchmark for real-world spatial understanding capabilities of multimodal ai models. https://huggingface.co/datasets/xai-org/RealworldQA. RealWorldQA is a benchmark designed for real-world understanding with 765 multiple-choice questions requiring recognition of details in high-resolution images.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in neural information processing systems*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Jinguo Zhu and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Appendix

A.1 Qualitative Examples

Refer to Fig. 5, with examples from each of the datasets, i.e VQAv2, RealWorldQA, CLEVR-Math, VAANI-H, JEE-T, VAANI-T, JEE-H

A.2 Evaluation Metrics

The RealWorldQA, VAANI-H/T subsets have only multiple-choice questions and JEE-T has multiple-choice or integer-answer questions. We use accuracy as the evaluation metric for all of these sets. We extract the answers using regex-based parsing, and report the overall accuracy across all the questions.

Hybrid Evaluation for VQA and CLEVR-Math: For VQAv2 and CLEVR-Math subsets, the answers are either a single word or short phrases. We adopt a hybrid evaluation strategy. We first evaluate a sample using exact match. Our exact match evaluation is built using the official VQA evaluation script (Goyal et al., 2017a), with the functionalities also extended to Hindi and Telugu. While exact match is strict and interpretable, it may penalize correct answers with minor surface-level variations (e.g., “yes” and “yes, it is”, synonyms, etc.). If exact match fails for a sample, we evaluate that sample using "gpt-4.1-2025-04-14" (OpenAI, 2025) as described in (Zheng et al., 2023). This two-step approach enables both high precision and flexibility, especially in cases in which answers may vary in form but not meaning. Its impact on scores is discussed in the appendix.

JEE-H Evaluation: The JEE-H Dataset has single correct MCQs, multiple correct MCQs, and numeric-type questions. We extend the scoring process used in (Arora et al., 2023). However, instead of the manual step they use, we replace with regex-based parsing to extract answers, and rule based processing to score each question. See appendix for details.

A.3 Ablation Study

We ablate prompting and other baselines using the Gemma-27B model on the Hindi subsets. We report comparisons for Direct Inference (Standard Prompting), Chain-of-Thought (Wei et al., 2022), and a Caption-Only (Deitke et al., 2025) baseline. Table 4 contains the results of ablation study.

Standard prompting provides a strong baseline performance across all datasets, especially on VAANI and VQA, indicating that the model benefits significantly from both modalities in relatively straightforward recognition tasks. In contrast, Chain-of-Thought (CoT) prompting notably improves performance on reasoning-heavy datasets. These improvements highlight the effectiveness of step-by-step reasoning prompts in tasks that require deeper cognitive processing. However, CoT does not benefit all tasks, with marginal gains or regression on VAANI and VQAv2, where reasoning depth is less critical. Overall, CoT benefits the English subset by 3.78 points in comparison to the benefit on Hindi at 0.79 points. This indicates a possible bias in the training data, with possibly more training done on English CoT data.

Next, to explore the need for visual embedding for QA, we evaluate a caption-only baseline. Here, the Gemma-27B model is asked to first generate a rich textual summary of the image. Then the generated caption and the original question are given to the model without the original image. This approach is inferior to the direct image-based inputs in all tasks.

A.4 Failure Analysis of Visual Language Models

We categorized the failures of GPT 4.1 on VAANI-T subset for both English and Telugu. We analyze results on VAANI-T as it sources images and QA from the Indian context and failures on this set point to common reasons for errors for current multimodal multilingual LLMs. We found 4 major categories of errors that appeared in both English and Telugu questions. They are: (1) Missing Indian context: Failed predictions in this category were due to missing knowledge specific to India needed to answer the question. (2) Visual Grounding error: Failures in this category were due to the model’s inability to associate objects in the picture with the question or options. (3) Visual Perception Failure: Errors in this category are due to the model failing to interpret the content of the image correctly. (4) Failure to Ground in Indian Context: Samples in this category accurately identified visual elements but they failed to analyze them in the appropriate socio-cultural context. Table 6 reports the percentages of the different categories on this dataset.

A.5 Related Work

Multimodal Instruction-Tuned Models. Previous work has shown that large language models can be extended to vision tasks by instruction tuning with multimodal data (Liu et al., 2023; Dai et al.,

Method	VQAv2		RealWorldQA		JEE-H		CLEVR-Math		VAANI-H		Ours-H	
	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En	Hi	En
Standard	**64.10**	**65.50**	54.38	61.04	19.66	17.58	43.50	53.70	**87.41**	82.01	53.81	55.97
CoT	60.50	62.70	**61.96**	**64.31**	**21.88**	**31.38**	**44.3**	**57.1**	84.34	**83.28**	**54.60**	**59.75**
Caption-Only	56.40	64.1	44.05	54.77	–	–	36.4	32.9	80.85	77.14	–	–

Table 4: Results on Gemma-27B across multiple datasets using different inference methods. **Bold** indicates the best performance, underline indicates the second best.

Benchmark	Indic Lang	QA Type	Image count	QA	Indic QA per lang	Human	Culturally Sourced
AyaVisionBench	**hin**	Chat	3.1k	3.1k	135	✗	✗
xMMMU	**hin**	MC	300	3k	291	✗	✗
xGQA	ben	OE	300	77.3k	9.7k	✗	✗
MTVQA	-	OE	8.8k	28.6k	-	✓	✗
M3Exam	-	MC	2.8k	12.3k	-	✓	✗
xChatBench	**hin**	Chat	400	400	50	✓	✗
Kaleidoscope	ben, **hin**, **tel**	MC	20.9k	20.9k	$\sim$ 800	✓	✗
XM100	ben, **hin**, **tel**	Caption	100	3.6k	100	✓	✓
MaRVL	tam	Caption	5.5k	5.7k	1.2k	✓	✓
MaXM	**hin**	OE	1.4k	2.1k	294	✓	✓
CVQA	ben, urd, tam, mar, **hin**, **tel**	MC	5.2k	10.4k	$\sim$ 300	✓	✓
Ours	**hin**, **tel**	OE, MC	5.1k	13.5k	$\sim 4k$	✓	✓

Table 5: Comparison of existing multilingual Visual Question Answering (VQA) benchmarks with ours.**"QA Type"** denotes the question-answering format (Chat = Multimodal chat; OE = open-ended VQA;MC = multiple-choice VQA; Caption = captioning/multi-image captioning); **"QA"** denotes the total question-answer pairs; **"Indic QA per lang"** denotes the question-answer pairs per Indic language; **"Human"** indicates human verified/annotated data; **"Culturally Sourced"** indicates culturally-sourced data. xChatBench, xMMMU and XM100 are part of PangeaBench (Yue et al., 2025). Ours (bottom row) is the only VQA dataset with human-verified annotations, diverse culturally sourced samples and a significant sample size per Indic language.

Error Category	English	Telugu
Lack of Knowledge about India	17%	16%
Visual Grounding Error	49%	50%
Visual Perception Failure	19%	18%
Failure to Ground in Indian Culture	15%	16%

Table 6: Error category distribution for GPT 4.1 on VAANI-T for English and Telugu. Percentages indicate the proportion of errors falling into each category.

2023; Li et al., 2023; Zhu et al., 2024). Many open weight Vision-Language Models such as Gemma3 (Team et al., 2025), Qwen2.5VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), Molmo (Deitke et al., 2025), Llama3.2 Vision (Meta Llama, 2025) and closed source models such as GPT-4.1 (OpenAI, 2025), Claude 3.7 (Anthropic, 2025), Gemini 2.5 (Google DeepMind, 2025)) have since been released, all leveraging some form of multimodal instruction tuning. Many of these recent models are multilingual, but the evaluation of their multilingual abilities is limited and is not reported on a consistent and reliable multimodal multilingual benchmark (Dash et al., 2025; Yue et al., 2025; Rasheed et al., 2025; Alam et al., 2025).

Multilingual Vision-Language Benchmarks. There have been a number of recent benchmarks which evaluate multi-lingual abilities and cultural robustness of VLMs. CVQA (Romero et al., 2024) contains 10.4k visual QA pairs across 31 languages, demonstrating large performance gaps on low-resource languages. MTVQA (Tang et al., 2024) provides text-centric VQA in 9 languages with human annotations. xGQA (Pfeiffer et al., 2022) automatically translates the GQA (Hudson and Manning, 2019) dataset into 7 languages while MaxM (Changpinyo et al., 2023) uses Machine Translation plus lightweight post-editing for 7 languages, with less than 300 samples per language. MarVL (Liu et al., 2021) focuses on culturally sourced reasoning by evaluating whether a caption about an image pair is true or false, but is not a Visual Question Answering dataset. Concurrently with our work, Kaleidoscope (Salazar et al., 2025) is a multi-

lingual multimodal benchmark with 800 VQA MCQs per Indic language sourced from regional examinations. However, Kaleidoscope has a narrow scope as it focuses only on exam-based questions and does not evaluate real-world understanding and practical reasoning. Multimodal benchmarks released along with their corresponding models include PangeaBench (Yue et al., 2025) (14 datasets in 47 languages), and AyaVisionBench (Dash et al., 2025) (9 tasks in 23 languages). However, these sets commonly have few manually verified samples per language, making the results less reliable for a particular target language. Table 5 provides an overview and compares them with our benchmark. A significant portion of our images per language (>1K) and our QA (>1K) are using culturally sourced, which is 5 to 20x larger than the culturally relevant subsets of previous datasets in this field.

Indic and Domain-Specific Datasets. Although there are dozens of English VQA resources (VQAv2 (Goyal et al., 2017a), CLEVR-Math (Lindström and Abraham, 2022), RealWorldQA (XAI.org, 2024)), there is a need for benchmarks in Indian languages. Previous works include (Singh et al., 2024; Arora et al., 2023). IndicGenBench (Singh et al., 2024) consists of 5 diverse tasks in 29 Indic languages but does not include multimodality. JEEBench (Arora et al., 2023) consists of questions from the highly competitive IIT JEE-Advanced exam, but does not include images-related questions. In contrast, our JEEset only includes questions with images. Our work fills the need for an accurate and diverse multimodal benchmark by providing ∼4k Visual Question Answering each in Hindi and Telugu.

A.6 Compute and Costs

We primarily ran all open-source models on H100 and A100 GPUs, rented via the Akash Console Network[2], incurring approximately 100 USD in compute costs. For proprietary models, we used APIs, leveraging Gemini's free tier and spending around 60 USD on OpenAI's APIs. The total expenditure amounts to approximately 160 USD.

A.7 Prompts

JEE-H COT Inference

```
eng_prompt = """You are an expert at solving
physics, chemistry and math questions that
involve both text and images. Analyze the text
and the associated image to answer the question
above.
First, think step-by-step and provide your
detailed reasoning. Explain how you interpret the
text and the image, the principles or formulas
you are using, and the intermediate calculations
or logical steps you take to arrive at the
solution.
After your reasoning, provide the final answer on
a new line.
Your entire response, including the reasoning and
the final answer, MUST end with the following
line exactly:
{ answer: }"""
```

JEE-H/T Standard Inference

```
eng_prompt = "You are an expert at solving
physics, chemistry and math questions that
involve both text and images.
Analyze the text and the associated image to
answer the question above. Provide only the final
answer, without any explanations or intermediate
steps. Your response MUST end with the following
line: { answer: }"
```

VQA LLM Eval

```
System_Prompt = "You are an expert judge
evaluating semantic similarity between two
answers.
Respond only with '1' for similar or '0' for not
similar."

User_Prompt = "Are the following two answers
semantically similar?

Answer 1: {a}
Answer 2: {b}

Respond with only '1' if they are similar in
meaning, or '0' if they are not similar."
```

CLEVR LLM Eval

```
System_Prompt = "You are an expert judge
evaluating semantic similarity between two
answers.
Respond only with '1' if they are semantically
similar, or '0' if they are not."

User_Prompt = "Are the following two {lang}
answers semantically similar?

Answer 1: {ans1}
Answer 2: {ans2}

Ignore minor differences in phrasing. Respond
with '1' if they express the same meaning, or '0'
otherwise."
```

A.8 Inference Strategies

This appendix details the various inference techniques employed for evaluating the multimodal models, along with the specific prompt setups used for English and Hindi. Below are setups in English.

Caption-Only Baseline

This technique involves a two-step process:
1) Dense Caption Generation: The model first

[2]https://akash.network/

generates a detailed, factual description (dense caption) of the provided image.

2) Question Answering with Caption: The model then answers the original question using *only* the generated dense caption as context, without access to the original image.

Prompt for Step 1 (Dense Caption Generation):

```
eng_prompt = Don't forget these rules:
1. Be Direct and Concise: Provide straightforward descriptions without adding interpretative or speculative elements.
2. Use Segmented Details: Break down details about different elements of an image into distinct sentences, focusing on one aspect at a time.
3. Maintain a Descriptive Focus: Prioritize purely visible elements of the image, avoiding conclusions or inferences.
4. Follow a Logical Structure: Begin with the central figure or subject and expand outward, detailing its appearance before addressing the surrounding setting.
5. Avoid Juxtaposition: Do not use comparison or contrast language; keep the description purely factual.
6. Incorporate Specificity: Mention age, gender, race, and specific brands or notable features when present, and clearly identify the medium if it's discernible.
When writing descriptions, prioritize clarity and direct observation over embellishment or interpretation. Write a detailed description of this image, do not forget about the texts on it if they exist. Also, do not forget to mention the type/style of the image. No bullet points. Start with the words, "This image displays:"
```

Prompt for Step 2 (Question Answering with Caption):

```
Context: {generated_caption}
Question: {question_text}
Answer:
```

Chain of Thought (CoT) Prompting

This technique guides the model to generate a step-by-step thought process (rationale) before arriving at the final answer. This is intended to improve reasoning capabilities for complex questions.

The image is provided along with language-specific system and user prompts.

English System Prompt:

```
When provided with an image and a question, generate a rationale first and then derive an answer.
Your rationale should include detailed visual elements in order to derive the answer.
```

English User Prompt:

```
Answer the question with following instruction:
1. Generate a rationale first and then derive an answer.
2. For your final answer, provide a Correct Option Number Only.
Question:
{question}
# Output Format #
<rationale>
### Answer: <your answer>
```

Standard (Direct Inference)

The model is provided with both the image and the question and is expected to generate a direct answer without explicit intermediate reasoning steps requested in the prompt.

The image and the question are provided to the model. The prompt typically instructs the model to answer directly, often in a specific format (e.g., option number for multiple-choice questions).

Example input (for a multiple-choice question):

```
Image: [Image Data]
Question: {question_text_with_options}
Instruction: Provide the Correct Option Number Only.
Answer:
```

(Note: This often shares the final answer formatting instruction with CoT, but without the explicit rationale generation step.)

JEE-H Scoring Rules

- For single correct Multi-Choice Questions (MCQs) and integer-type questions, a binary score was assigned: 1 if the model answered with the correct option/integer and 0 otherwise.
- For multi-correct MCQs, a score of 1 is given when the generated response consists of all and only the correct options. If any of the wrong option is chosen, the response is given a score of 0. If the response contains no incorrect option and some of the correct options, a score of 0.25 is given for each of the correct options.
- For numeric-type questions, answers in the range of ±0.01 with the gold answer are given a score of 1, and 0 otherwise.
- In Chain-of-Thought (CoT) inferencing, if a Hindi response included reasoning in English, 0 was given irrespective of the answer.
- In standard zero-shot inference, where the model was explicitly instructed to return only a direct answer, responses containing any explanation or reasoning were also scored 0, irrespective of the answer.

Impact of LLM Judge on Evaluation Scores

As detailed in Section A.2, our hybrid evaluation strategy for CLEVR-Math and VQA incorporates an LLM judge to re-evaluate answers initially marked incorrect by the exact match protocol. This approach revealed a notable trend: the score increase attributed to the LLM judge was considerably more for the responses in Hindi and Telugu as compared to English. For instance, when evaluating the VQA dataset, the average scores for English responses increased by $\approx 4\%$ after LLM judging. In contrast, Hindi scores on the same dataset increased by $\approx 12\%$ after the initial exact match scores. This disparity suggests that the models demonstrate a greater proficiency in adhering to strict answer formatting conventions in English, while responses in Hindi and Telugu, though often semantically correct, are more frequently penalized by the exact match criteria due to variations in phrasing or formatting. The LLM judge effectively mitigates this by recognizing a broader range of correct expressions.

A.9 Licenses

The datasets used in our benchmark are released under the following licenses:

- VQAv2, VAANI and ClevrMath are released under Creative Commons Attribution 4.0 International (CC BY 4.0) license.
- RealWorldQA is released under Creative Commons Attribution No Derivatives 4.0.
- JEE-Vision is built from publicly available JEE questions. We will be releasing links under CC BY-NC 4.0 license along with the QA pair. The exam papers are distributed for the general public (https://www.jeeadv.ac.in/archive.html) used by various publishers and institutes for commercial purposes already. When collecting this data, we followed the steps by previous works on JEE Bench (Arora et al., 2023).

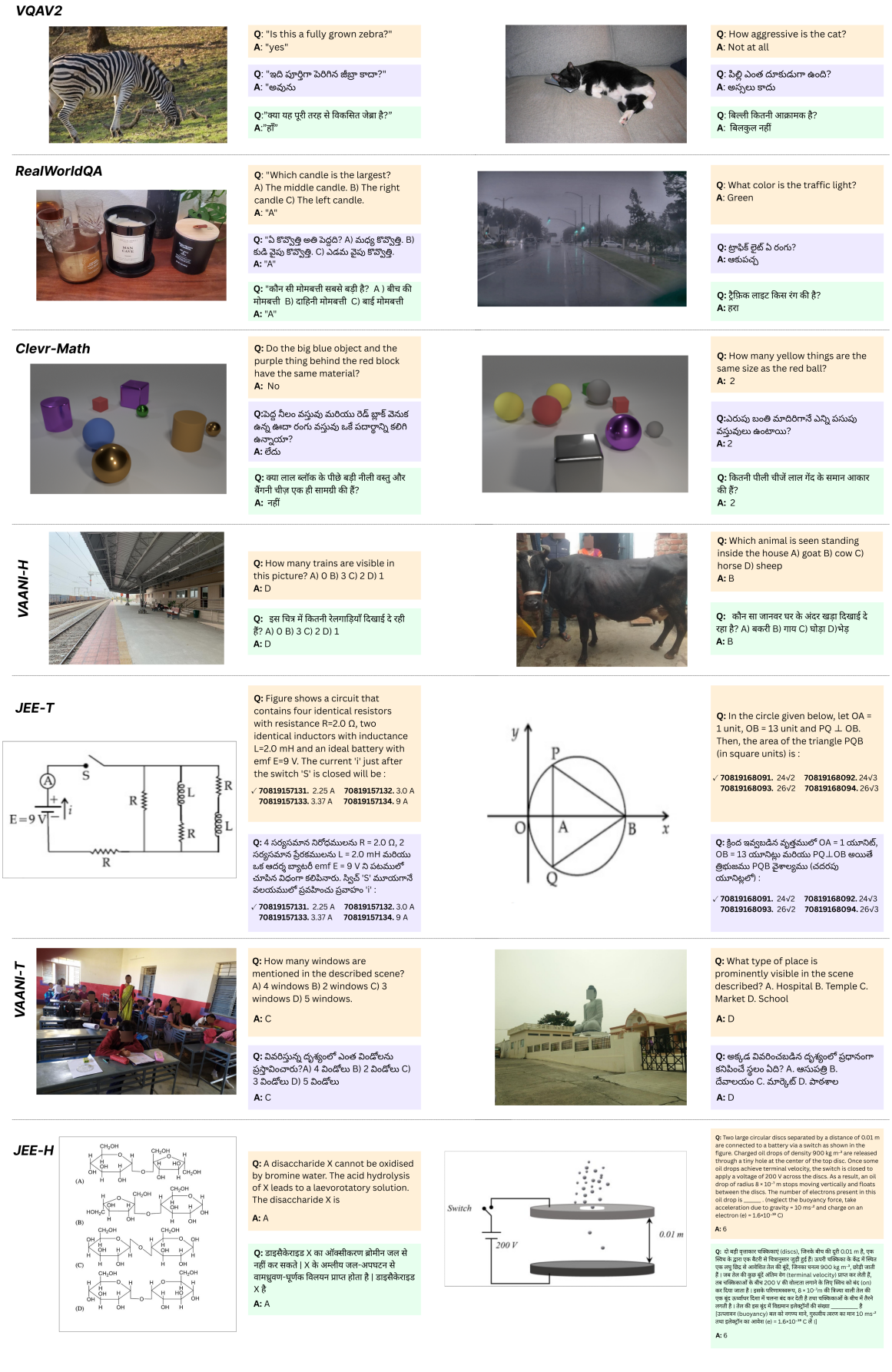

Figure 5: Qualitative Examples from ***VQAv2, RealWorldQA, CLEVR-Math, VAANI-H, JEE-T, VAANI-T, JEE-H***