arXiv:2505.15075v1 [cs.CL] 21 May 2025

Traveling Across Languages: Benchmarking Cross-Lingual Consistency in Multimodal LLMs

Hao Wang[1*] **Pinzhi Huang**[2] **Jihan Yang**[2] **Saining Xie**[2] **Daisuke Kawahara**[13]
[1]Waseda University [2]New York University [3]NII LLMC

https://github.com/nlp-waseda/traveling-across-languages

Abstract

The rapid evolution of multimodal large language models (MLLMs) has significantly enhanced their real-world applications. However, achieving consistent performance across languages, especially when integrating cultural knowledge, remains a significant challenge. To better assess this issue, we introduce two new benchmarks: **KnowRecall** and **VisRecall**, which evaluate cross-lingual consistency in MLLMs. KnowRecall is a visual question answering benchmark designed to measure factual knowledge consistency in 15 languages, focusing on cultural and historical questions about global landmarks. VisRecall assesses visual memory consistency by asking models to describe landmark appearances in 9 languages without access to images. Experimental results reveal that state-of-the-art MLLMs, including proprietary ones, still struggle to achieve cross-lingual consistency. This underscores the need for more robust approaches that produce truly multilingual and culturally aware models.

1 Introduction

Multimodal large language models (MLLMs) have recently undergone rapid progress, giving rise to a wide range of practical applications (Zhang et al., 2024). While the computer vision community has extensively studied their vision perception capabilities (Tong et al., 2024; Fu et al., 2024), the multilingual dimension of MLLMs remains relatively underexplored. In particular, their performance often deteriorates when applied to languages with limited resources or distinct cultural contexts.

To bridge the gap, recent studies have focused on developing multimodal culture understanding benchmarks (Liu et al., 2021; Nayak et al., 2024; Romero et al., 2024; Vayani et al., 2024) and training more powerful multilingual MLLMs (Chen et al., 2023; Yue et al., 2025; Geigle et al., 2025; Dash et al., 2025). However, current models still exhibit varying performance across languages, falling short of the ideal goal—providing consistent responses regardless of input language. While some recent studies have examined cross-lingual consistency in text-only LLMs (Qi et al., 2023; Gao et al., 2024; Huang et al., 2024b; Wang et al., 2025), no research has yet explored this issue on MLLMs, even though ensuring consistent behavior across languages is crucial for real-world applications.

To address this issue, we propose two novel benchmarks: KnowRecall and VisRecall, designed to evaluate cross-lingual consistency in multilingual MLLMs under a traveling scenario. KnowRecall is a visual question answering (VQA) benchmark that assesses the consistency of factual knowledge across 15 languages, focusing on cultural and historical questions about global landmarks. Meanwhile, VisRecall evaluates the consistency of visual generation by instructing models to describe the appearance of landmarks in 9 languages without direct visual input during inference.

Through extensive experiments on state-of-the-art open-weight and proprietary MLLMs, we observe persistent challenges in multilingual alignment. Particularly, performance consistently declines from English to local languages of the corresponding landmarks, and drops even further in other foreign languages. While models show high consistency scores within related language families, such as Romance languages, their performance still lags in lower-resource settings. We also find that inference-time reasoning yields notable improvements, implying that leveraging models' reasoning ability (Snell et al., 2024; DeepSeek-AI et al., 2025) could be a promising direction for tackling language constraints. Moreover, in the VisRecall task, models that have directly "seen" these landmarks during multimodal training fail to effectively leverage their visual memory for multilingual description generation, indicating a fundamental

[*]Work conducted during a visit to NYU.

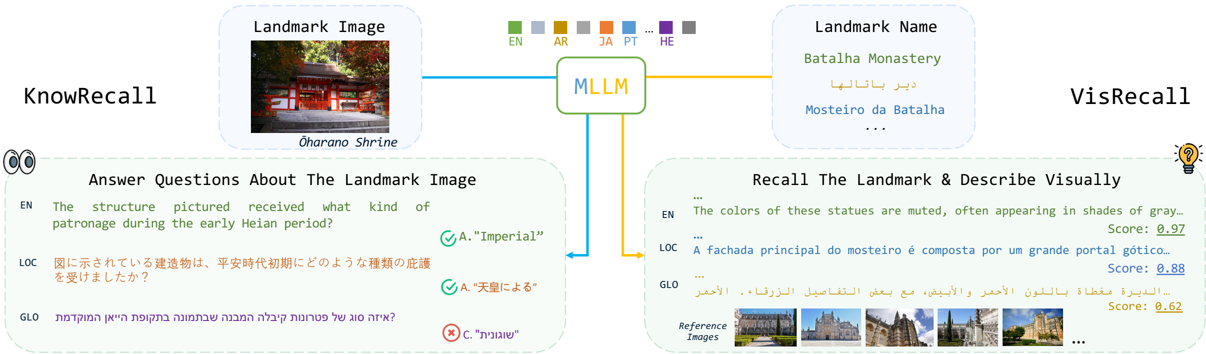

Figure 1: Illustrations of KnowRecall and VisRecall. KnowRecall evaluates the cross-lingual consistency of factual knowledge in MLLMs using a VQA setup, where the model answers questions about a given landmark image in 15 languages. VisRecall measures the cross-lingual consistency of visual memory by assessing the quality of landmark descriptions generated in 9 languages, using CLIPScore for evaluation.

disconnect between current multimodal training paradigms and human-like visual cognition.

2 KnowRecall

Imagine a French tourist visiting Tokyo Tower, snapping a photo and asking an MLLM about the tower's height. Naturally, they would expect a correct response in their native language. However, if the model provides the right answer in Japanese but fails to do so in French, it illustrates a critical real-world limitation. We introduce KnowRecall, a multilingual VQA benchmark that evaluates cross-lingual consistency of factual knowledge in MLLMs. Unlike existing multilingual culture understanding benchmarks (e.g., Romero et al., 2024) which include questions only in English and the local language, our dataset offers 3,000 multiple-choice questions on 1,500 global landmarks, each available in 15 languages. This breadth facilitates a comprehensive assessment of cross-lingual consistency across diverse linguistic contexts.

Dataset Creation We selected 15 target languages based on speaker population and geographic diversity. We sampled 100 landmarks for each language from the Google Landmarks Dataset v2 (GLDv2, Weyand et al., 2020), selecting only those located in countries with a single official language (e.g., Canada was excluded due to its dual official languages). For each landmark, we manually chose a single representative image to maintain data quality. For VQA generation, we adapted the framework from Su et al. (2024), leveraging Gemini-1.5-Pro to generate two questions per landmark based on the associated image and its English Wikipedia page. We then used Gemini to translate these questions into the remaining 14 languages.

Evaluation Metrics Following Antol et al. (2015); Romero et al. (2024), we use accuracy to measure model performance. Instead of evaluating performance in solely English and multilingual settings, we introduce a new evaluation scheme with three distinct settings: **EN** (questions are in English), **LOC** (questions are in the local language of each landmark), and **GLO** (the average performance across all languages except English and the local language). The GLO setting better aligns with real-world inbound tourism needs, offering a novel perspective for evaluating multilingual MLLMs.

Inspired by Jiang et al. (2020); Gao et al. (2024), we measure cross-lingual consistency using the ratio of correct predictions shared between two languages. Let n_x and n_y denote the number of correct answers in languages x and y, respectively, with n_{xy} representing the number of answers correct in both, we define consistency as:

$$\text{Consistency}_{\text{K}}(x, y) = \frac{1}{2}\left(\frac{n_{xy}}{n_x} + \frac{n_{xy}}{n_y}\right) \quad (1)$$

We compute the consistency of each local language with the other 14 languages and obtain the final score by averaging across all language pairs.

3 VisRecall

The tourist finished the journey and came back to France, eager to share the places they visited with their friends. When portraying these experiences, the visual information they convey is inherently independent of language, meaning that descriptions created in different languages should ideally be highly similar. This concept extends to MLLMs as well. While a model may demonstrate decent consistency in VQA tasks, any inconsistency in generation tasks would lead to a biased user experience (i.e., a knowing *vs* saying distinction Orgad et al., 2024; Brinkmann et al., 2025). To assess the cross-lingual consistency of "visual memory" in MLLMs, we introduce VisRecall, a multilingual benchmark designed to evaluate visual description generation across 450 landmarks in 9 languages.

Dataset Creation Due to current MLLMs' limited generation capabilities in low-resource languages, we restrict VisRecall to 9 target languages for more reliable evaluation (see Appendix C for details). For each language, we sampled 50 relatively well-known landmarks from GLDv2, ensuring that all 9 languages have corresponding Wikipedia pages for each landmark. The task input is the landmark's name in each language, and the output is the description generated by the models.

Evaluation Metrics A landmark's appearance description can vary depending on factors such as orientation, viewing angle, and weather conditions, making it challenging even for humans to establish a definitive ground truth. To address this, we leverage CLIPScore (Hessel et al., 2022) for reference-free evaluation. We selected up to 20 images per landmark from GLDv2 and compute the CLIPScore between the generated description and each image. For non-English descriptions, we first translate them into English using Gemini-1.5-Pro before evaluation. The final score for each landmark-language pair is then calculated by averaging the CLIPScore across all selected images.

We define consistency for VisRecall as:

$$\text{Consistency}_{\text{V}}(x,y) = \frac{1}{2}\left(\frac{S}{\sum_i s_x^{(i)}} + \frac{S}{\sum_i s_y^{(i)}}\right) \qquad (2)$$

where $S = \sum_i \min(s_x^{(i)}, s_y^{(i)})$, with $s_x^{(i)}$ and $s_y^{(i)}$ as the ith landmark's CLIPScore in language x and y.

4 Experiments and Results

We select a range of MLLMs as baselines to evaluate performance on KnowRecall and VisRecall. For KnowRecall, we evaluate the models in a zero-shot manner, instructing them to directly output the correct answer option. For VisRecall, given that language models are highly sensitive to subtle variations in prompts (Sclar et al., 2024; Yin et al., 2024), we design two prompt templates per language with minimal cross-linguistic differences. The full list of

Model	EN	LOC	GLO	Consistency
LLaVA-1.5-7B	43.8	38.7	35.1	58.3
LLaVA-OV-7B	51.1	45.7	42.5	71.3
Pangea-7B	54.2	51.4	48.6	77.9
Qwen2.5-VL-7B-IT	<u>56.6</u>	<u>55.2</u>	<u>51.0</u>	<u>80.9</u>
Cambrian-8B	46.3	43.1	39.8	65.5
InternVL2.5-8B	51.2	44.7	41.2	64.8
Llama-3.2-11B-V-IT	50.2	48.1	46.9	73.8
Gemini-1.5-Pro	63.9	61.4	57.2	84.0
Gemini-2.0-Flash	64.5	65.1	59.0	**86.3**
GPT-4o	**68.3**	**69.2**	**64.4**	85.9

Table 1: Performance on KnowRecall. The best-performing open-weight model is <u>underlined</u> and the best proprietary model is in **bold**.

Model	EN	LOC	GLO	Consist.	LangAd (%)
Llama-3-8B-IT[T]	81.9	79.1	75.0	95.8	30.9
Cambrian-8B	76.8	73.4	69.7	93.8	99.7
InternLM2.5-7B-Chat[T]	81.5	78.0	74.2	95.4	93.1
InternVL2.5-8B	79.8	76.7	73.6	95.5	99.8
Qwen2-7B-IT[T]	<u>82.7</u>	<u>80.1</u>	<u>77.3</u>	<u>96.6</u>	99.9
Pangea-7B	79.4	77.1	74.5	96.2	100.0
Qwen2.5-7B-IT[T]	78.8	78.5	75.5	96.0	98.9
Qwen2.5-VL-7B-IT	80.3	78.9	75.9	96.4	99.9
Gemini-1.5-Pro	74.9	73.8	72.1	96.1	100.0
Gemini-2.0-Flash	75.7	74.9	73.3	96.3	100.0
GPT-4o	**80.1**	**80.4**	**79.3**	**97.5**	100.0

[T]Text-only LLMs.
···: Each pair of models separated by a dotted line consists of a LLM back-born and an MLLM trained on top of it.

Table 2: Performance on VisRecall, where LangAd measures the proportion of outputs adhering the prompt's language, detecting using Lingua (Stahl). Notably, while Llama-3-8B-IT scores high, it often fails to follow the prompt language, defaulting to English instead.

prompts is provided in Appendix H.2. During evaluation, we compute the final score as the average of the results from both prompts. Since VisRecall does not require images as input, we also select several text-only LLMs to compare whether MLLMs, trained on a large volume of caption data, exhibit a stronger visual memory of landmarks.

We show the KnowRecall results in Table 1. Overall, the models achieve their best performance in the EN setting, reflecting the predominance of English in their pre-training. In the LOC setting, open-weight models show a slight decline in performance compared to the EN setting, while proprietary models maintain comparable results. Notably, all models, including proprietary ones, consistently perform worst in the GLO setting. A similar trend is observed in the VisRecall results (Table 2), where performance follows the pattern EN $\geq$ LOC $>$

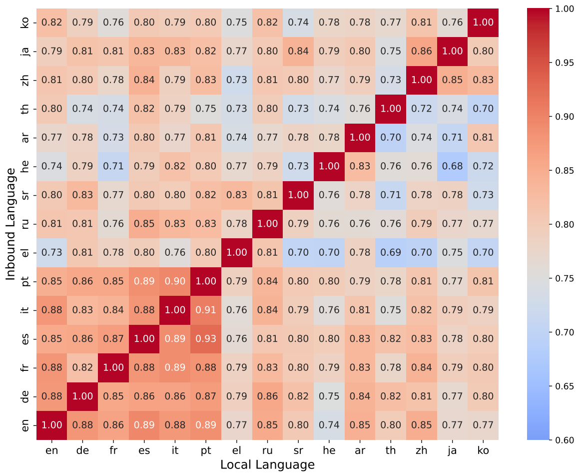

Figure 2: Consistency score matrix of Qwen2.5-VL-7B-IT on KnowRecall. Each cell (x, y) denotes the score between language x and y, based on questions about landmarks in regions where x is the local language.

GLO. This highlights the substantial gap in multilingual capabilities among current MLLMs and underscores the potential risks of deploying these models in real-world multilingual applications.

5 Discussion

5.1 Consistency in related language families

As shown in Figure 2, while Qwen2.5-VL-7B-IT achieves the highest consistency score among open-weight models, consistency varies across languages. We observe high consistency scores within related language families, such as Germanic (English and German) and Romance (French, Spanish, Italian, and Portuguese). Similarly, Chinese and Japanese show strong consistency, likely due to their shared character systems. In contrast, comparable lower-resource languages, such as Greek and Hebrew, still exhibit relatively low consistency, suggesting barriers to effective multilingual alignment.

5.2 Inference-Time Reasoning

To evaluate the impact of inference-time reasoning, we design a structured chain-of-thought (CoT, Wei et al., 2023) prompt (see Appendix H.1) for KnowRecall. This prompt systematically guides the model through three steps: (1) recognizing the landmark; (2) translating the question into the local language or English; and (3) reasoning through to produce a final answer. As shown in Table 3, Gemini-2.0-Flash and GPT-4o achieve notable gains in accuracy and consistency, demonstrating the benefits of inference-time reasoning.

Model	EN	LOC	GLO	Consistency
Gemini-2.0-Flash	64.5	65.1	59.0	86.3
+ Structured CoT (LOC)	68.6	67.9	66.3	88.9
+ Structured CoT (EN)	68.1	67.9	65.5	88.2
GPT-4o	68.3	69.2	64.4	85.9
+ Structured CoT (LOC)	72.3	72.6	68.8	89.3
+ Structured CoT (EN)	73.1	71.6	69.0	89.8

Table 3: Performance boost through inference-time reasoning on KnowRecall. Structure CoT (LOC) translates questions into the local language, whereas Structure CoT (EN) translates them into English.

Although this approach does not fully address cross-lingual alignment—largely bypassing the issue by leveraging geographical knowledge and translation, it illustrates a promising direction for harnessing language models' reasoning abilities to overcome linguistic constraints.

5.3 The effect of multimodal training

In Table 2, each pair of models separated by a dotted line consists of a LLM back-born and an MLLM trained on top of it. As the landmarks are relatively famous, they are expected to occur repeatedly during the MLLMs' multimodal training. In other words, these models have directly "seen" the landmarks, suggesting they should possess strong visual knowledge of their appearances. However, interestingly, with the exception of the Qwen2.5 pair, all base LLMs outperform their corresponding MLLMs in both CLIPScore and consistency. This indicates that MLLMs may struggle to fully leverage the visual knowledge acquired during multimodal training, likely due to the significant differences in prompting paradigms. Nevertheless, the ability to generalize such information is crucial for real-world applications such as robotics and autonomous driving. We argue that VisRecall serves as a suitable assessment standard for this challenge.

6 Conclusion

In this paper, we introduced KnowRecall and VisRecall to systematically evaluate cross-lingual consistency in MLLMs. Our experiments revealed notable gaps across different languages, especially in low-resource settings, highlighting the need for more robust alignment. Key insights from our paper include: (1) Models achieve higher consistency within related language families; (2) Structured chain-of-thought prompting improves consistency by leveraging reasoning and translation capabilities; (3) Text-only models often outperform multimodal ones, indicating difficulties in integrating visual memory. We hope these findings, along with the proposed benchmarks, will catalyze further research toward developing truly multilingual and culturally attuned MLLMs.

Limitations

We constructed the KnowRecall dataset using Gemini for translation. Given the extensive number of language variants and the large volume of VQA questions, it was impractical to double-verify every translation. Consequently, some translation errors may be present. We provide further discussion on Gemini's translation quality in Appendix F.

During evaluation on VisRecall, we observe that shorter outputs tend to result in lower CLIPScore, which make it challenging to compare absolute scores across different models. For instance, two Gemini models, despite their strong multimodal and multilingual capabilities, exhibit unusually low scores (Table 2). Given this limitation, we recommend prioritizing consistency scores and comparing CLIPScore only within related model families while using VisRecall. We plan to improve this in future iterations of our benchmark.

Acknowledgements

We thank Shusheng Yang, Boyang Zheng, Ziteng Wang, Zihao Yang, Shuhei Kurita and Haiyue Song for their helpful discussions and feedback.

This work was supported by the Gemma Academic Program for JP/KR 2024, and the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology. H.W. thanks the financial support provided by the Future Robotics Organization at Waseda University for the visit to NYU.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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. *Preprint*, arXiv:2502.13923.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. *Preprint*, arXiv:2501.06346.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, and 24 others. 2023. Pali-x: On scaling up a multilingual vision and language model. *Preprint*, arXiv:2305.18565.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

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. *Preprint*, arXiv:2505.08751.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. *Preprint*, arXiv:2404.04659.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and

Goran Glavaš. 2025. Centurio: On drivers of multilingual ability of large vision-language model. *Preprint*, arXiv:2501.05122.

Gemini-Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

Kaichen Huang, Jiahao Huo, Yibo Yan, Kun Wang, Yutao Yue, and Xuming Hu. 2024a. Miner: Mining the underlying pattern of modality-specific neurons in multimodal large language models. *Preprint*, arXiv:2410.04819.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 2024b. 1+1>2: Can large language models serve as cross-lingual knowledge aggregators? *Preprint*, arXiv:2406.14721.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *Preprint*, arXiv:2412.08802.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

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, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 10650–10666. Association for Computational Linguistics.

Qwen, 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, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

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. *Preprint*, arXiv:2406.05967.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep nlp models: A survey. *Preprint*, arXiv:2108.13138.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Peter M. Stahl. lingua-py: The most accurate natural language detection library for python. Version 2.0.2.

Xin Su, Man Luo, Kris W Pan, Tien Pei Chou, Vasudev Lal, and Phillip Howard. 2024. Sk-vqa: Synthetic knowledge generation at scale for training context-augmented multimodal llms. *Preprint*, arXiv:2406.19593.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Preprint*, arXiv:2406.16860.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Yumeng Wang, Zhiyuan Fan, Qingyun Wang, May Fung, and Heng Ji. 2025. Calm: Unleashing the cross-lingual self-aligning ability of language model question answering. *Preprint*, arXiv:2501.18457.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

T. Weyand, A. Araujo, B. Cao, and J. Sim. 2020. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. *Preprint*, arXiv:2402.14531.

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. *Preprint*, arXiv:2410.16153.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Preprint*, arXiv:2401.13601.

A Related Work

Cross-lingual Consistency Qi et al. (2023) examined the cross-lingual consistency of factual knowledge in multilingual pre-trained language models, finding that while larger models improve factual accuracy, they do not enhance consistency. Similarly, Gao et al. (2024) explored the impact of multilingual pre-training and instruction tuning on alignment, highlighting that their effectiveness depends on the chosen strategy—where continued pre-training can benefit target languages but may come at the cost of others. To address language performance disparities, Huang et al. (2024b) proposed a framework that aggregates knowledge across languages, demonstrating improvements in multilingual LLM performance. Wang et al. (2025) introduced a DPO-based (Rafailov et al., 2024) method to enhance knowledge consistency in multilingual LLMs, showing its effectiveness on medical and commonsense QA datasets.

B Potential Solutions

Besides inference-time reasoning explored in Section 5.2, we also plan to experiment with several additional directions to enhance benchmark performance in future work. For instance, DPO (Rafailov et al., 2024) and GRPO (Shao et al., 2024) are promising techniques for bridging the performance gap between English and other languages. Moreover, neuron-level interpretation and control (Sajjad et al., 2022) has gained popularity as a research direction. Previous studies demonstrated the existence of language-specific neurons controlling output languages (Kojima et al., 2024), as well as modality-specific neurons controlling modality perception (Huang et al., 2024a) in LLMs. Inspired by these findings, identifying and steering *region-specific* neurons in MLLMs (e.g., Japan-specific neurons) might enable leveraging visual inputs to further narrow the cross-lingual performance gap.

C The List of Languages

We list the 15 languages selected in KnowRecall and the 9 languages selected in VisRecall in Table 4.

Name	ISO-639	KnowRecall	VisRecall
Arabic	ar	✓	✓
Chinese	zh	✓	✓
English	en	✓	✓
French	fr	✓	✓
German	de	✓	✓
Greek	el	✓	✗
Hebrew	he	✓	✗
Italian	it	✓	✓
Japanese	ja	✓	✓
Korean	ko	✓	✗
Portuguese	pt	✓	✓
Russian	ru	✓	✗
Serbian	sr	✓	✗
Spanish	es	✓	✓
Thai	th	✓	✗

Table 4: Languages and their corresponding language codes selected in KnowRecall and VisRecall datasets.

D Model Cards

We list the models used in the paper in this section.

D.1 Proprietary models

- Gemini-1.5-Pro (Gemini-Team et al., 2024):`gemini-1.5-pro-002`
- Gemini-2.0-Flash: `gemini-2.0-flash-001`
- GPT-4o (OpenAI et al., 2024): `gpt-4o-2024-11-20`

D.2 Open-weight models

- LLaVA-1.5-7B (Liu et al., 2024):liuhaotian/llava-v1.5-7b
- LLaVA-OV-7B (Li et al., 2024):lmms-lab/llava-onevision-qwen2-7b-ov
- Pangea-7B (Yue et al., 2025):neulab/Pangea-7B
- Qwen2-7B-IT (Yang et al., 2024):Qwen/Qwen2-7B-Instruct
- Qwen2.5-7B-IT (Qwen et al., 2025):Qwen/Qwen2.5-7B-Instruct
- Qwen2.5-VL-7B-IT (Bai et al., 2025): Qwen/Qwen2.5-VL-7B-Instruct
- Cambrian-8B (Tong et al., 2024): nyu-visionx/cambrian-8b
- InternLM2.5-7B-Chat (Cai et al., 2024): internlm/internlm2_5-7b-chat
- InternVL2.5-8B (Chen et al., 2025): OpenGVLab/InternVL2_5-8B
- Llama-3-8B-IT (Grattafiori et al., 2024): meta-llama/Meta-Llama-3-8B-Instruct
- Llama-3.2-11B-V-IT (Grattafiori et al., 2024): meta-llama/Llama-3.2-11B-Vision-Instruct

D.3 CLIP model

- Jina CLIP v2 (Koukounas et al., 2024): jinaai/jina-clip-v2

E Qualitative Examples

E.1 KnowRecall with structured CoT

Figure 3 and Figure 4 present outputs from Gemini-2.0-Flash using structured CoT prompts.

w/o structured CoT prompt | w/ structured CoT prompt (LOC) | w/ structured CoT prompt (EN)

Question in Korean:
이미지에 보이는 건물은 제2차 세계 대전 중 어떤 은행이 점거했습니까?
A: 중국은행
B: 요코하마 정금 은행 ✅
C: 도쿄 은행
D: 도이치 은행

B

The image shows a building with a dome and a clock tower, which are landmarks of the Bund in Shanghai. Therefore, the location is Shanghai, China.

Translation into Chinese:
问题：图片中显示的建筑在第二次世界大 战期间被哪家银行占据？
选项：A. 中国银行 B. 横滨正金银行 C. 东京银行 D. 德意志银行

Reasoning in Chinese:
这张图片是上海外滩汇丰银行大楼。二战期间，该建筑曾被日本的横 滨正金银行占用。其他选项的银行，中国银行是中国银行自己的建筑，东京银行是日本的，但是没有占用这个建筑，德意志 银行是德国的，也没有占用 这个建筑。因此，正确答案是横滨正金银行。

Final Answer:
{"answer": "B"}

1. The image shows a large, stately building with a dome and columns. The architecture and the presence of a clock tower in the background suggest that it is located in Shanghai, likely on The Bund. The building in question is the former HSBC building.

2. Translation of the question:
Which bank occupied the building shown in the image during World War II?
Translation of the options:
A. Bank of China B. Yokohama Specie Bank C. Bank of Tokyo D. Deutsche Bank

3. The image shows the former HSBC building on The Bund in Shanghai. During World War II, Shanghai was occupied by the Japanese. Of the options provided, the Yokohama Specie Bank (Yokohama Shokin Ginko) was a Japanese bank. HSBC's assets were frozen and the bank placed under Japanese control.

4. {"answer": "B"}

Figure 3: A set of example outputs from Gemini-2.0-Flash on KnowRecall. The question is asked in Korean, while the local language is Chinese.

w/o structured CoT prompt	w/ structured CoT prompt (LOC)	w/ structured CoT prompt (EN)

Question in German:
Das Gebäude im Bild ist das älteste erhaltene Gebäude des Tempels. In welchem Architekturstil ist es erbaut?
A: Kasuga-zukuri
B: Nagare-zukuri
C: Irimoyazukuri ✅
D: Taisha-zukuri

Taisha-zukuri

Here's the breakdown:

1. Location Analysis:
The architecture and surrounding environment strongly suggest the building is located in Japan. The temple architecture with the specific roof style seen in the image is characteristic of Japanese shrines.

2. Translation to Japanese:
Question: 写真の建物は、その寺で最も古い現存する建物です。それはどのような建築様式で建てられていますか？
Options: A. 春日造 B. 流造 C. 入母屋造 D. 大社造

3. Reasoning in Japanese:
建物の屋根の形をよく見ると、寄棟造と切妻造が組み合わさった入母屋屋根であることがわかります。したがって、この建物は入母屋造の建築様式で建てられています。

4. JSON Output:
{"answer": "C"}

1. Based on the image, the architecture style is likely from Japan. The building is surrounded by trees, a typical setup for Shinto shrines.

2. Translation:
Question: The building in the picture is the oldest remaining building of the temple. In which architectural style is it built?
A. Kasuga-zukuri B. Nagare-zukuri C. Irimoyazukuri D. Taisha-zukuri

3. Reasoning:
Looking at the image, the building has a gabled roof, but it's not asymmetrical like in Nagare-zukuri. It also doesn't seem to have the distinctive features of Taisha-zukuri or Irimoyazukuri. However, the building closely resembles the Kasuga-zukuri style with its simple rectangular form and gabled roof.

4. Final Answer:
{"answer": "A"}

Figure 4: A set of example outputs from Gemini-2.0-Flash on KnowRecall. The question is asked in German, while the local language is Japanese. Part of the reasoning process has been omitted for clarity.

E.2 VisRecall

We present outputs from Qwen2.5-VL-7B-IT in Figure 5.

Batalha Monastery

5-20 Filtered Landmark Images

Describe Visually in Portuguese (LOC)

1
A fachada principal do Mosteiro da Batalha é imponente, com uma torre central alta e esguia que se eleva acima das outras estruturas. A torre é feita de pedra cinza, com detalhes em branco nas janelas e no topo. As paredes do mosteiro são feitas de pedra cinza, com algumas partes em pedra vermelha. O telhado é feito de telhas de barro, com uma cor marrom-avermelhada. O ambiente ao redor do mosteiro é verde e exuberante, com árvores altas e arbustos densos. Há um pequeno riacho que corre ao lado do mosteiro, com água cristalina e correnteza suave. O céu está nublado, com nuvens cinzas e pesadas que cobrem o sol. A luz do dia é fraca, criando sombras longas e profundas nas paredes do mosteiro.

2
Aqui está uma descrição visual do Mosteiro da Batalha:
O Mosteiro da Batalha é uma construção monumental de pedra, com suas paredes grossas e robustas que parecem ter sido moldadas pelo próprio tempo. A estrutura é composta principalmente de blocos de arenito, um material que ganha uma tonalidade amarela quando exposto à luz do sol, criando um contraste atraente com as sombras escuras que se projetam sobre ele.
As torres do mosteiro são altas e imponentes, com telhados de telha vermelha que contrastam com o céu azul. As janelas são estreitas e retangulares, permitindo que a luz entre em pequenas quantidades, criando um ambiente misterioso e quase sagrado.
O interior do mosteiro é uma mistura de luz e sombra, com a luz do sol brilhando através das janelas, iluminando os detalhes da arquitetura. O chão é coberto por ladrilhos de pedra, que têm uma textura rugosa e desgastada, refletindo a passagem do tempo.
A vegetação circunda o mosteiro, com árvores e arbustos que crescem ao redor, criando uma sensação de harmonia com a natureza. O verde das folhas se mistura com as cores mais quentes da pedra, criando uma paleta de cores vibrante e atraente.
A atmosfera do mosteiro é de serenidade e tranquilidade, mas também de poder e majestade. É como se cada pedra tivesse uma história a contar, cada janela tivesse um olhar para o passado, e cada árvore tivesse uma conexão com a natureza. É um lugar que convida o visitante a parar, a respirar, a apreciar a beleza do momento presente.

Description – (PT translated to EN)

1
The Batalha Monastery is an imposing architectural structure, built primarily with white limestone, which contrasts sharply with the green of the surrounding trees. The main façade of the monastery features a large Gothic portal, with arches supported by slender and detailed columns, and a tympanum decorated with reliefs. The building has several towers, some with red-tiled roofs, while others are covered with ceramic tiles. The windows are narrow and rectangular, with stone frames that stand out against the white of the limestone. The interior of the monastery is vast and houses several rooms, including the main chapel, which is notable for its height and width, with stone walls and a wooden ceiling. The floor is covered with stone tiles, and the walls are lined with colorful tiles. Natural light enters through the windows and openings in the walls, creating a somber and mysterious atmosphere. Around the monastery, there is a vast garden with leafy trees and shrubs, which provide a counterpoint to the rigidity of the building's walls. The site is surrounded by a high stone fence, which protects the monastery and marks its presence in the landscape. In addition, there is a small stream running near the monastery, adding a touch of nature to the setting.

2
Here is a visual description of the Batalha Monastery: The Batalha Monastery is a monumental stone construction, with its thick and robust walls that seem to have been shaped by time itself. The structure is primarily composed of sandstone blocks, a material that gains a yellowish hue when exposed to sunlight, creating an attractive contrast with the dark shadows cast upon it. The towers of the monastery are tall and imposing, with red-tiled roofs that stand out against the blue sky. The windows are narrow and rectangular, allowing light to enter in small amounts, creating a mysterious and almost sacred atmosphere. The interior of the monastery is a mixture of light and shadow, with sunlight streaming through the windows, illuminating the architectural details. The floor is covered with stone tiles, which have a rough and worn texture, reflecting the passage of time. Vegetation surrounds the monastery, with trees and shrubs growing around it, creating a sense of harmony with nature. The green of the leaves blends with the warmer tones of the stone, creating a vibrant and attractive color palette. The atmosphere of the monastery is one of serenity and tranquility, but also of power and majesty. It feels as if each stone has a story to tell, each window offers a glimpse into the past, and each tree has a connection to nature. It is a place that invites the visitor to pause, to breathe, to appreciate the beauty of the present moment.

PT CLIPScore

1: 0.879	2: 0.895

Describe Visually in Arabic (GLO)

1
دير باتالها هو بناء ضخم يقع في منطقة باتالها في ولاية مهاراشترا الهندية. يُعرف الدير بتصميمه الفريد الذي يجمع بين العناصر الهندوسية والمعمارية الإسلامية.
يتميز الدير بواجهته الرئيسية التي تتألف من أربعة أعمدة ضخمة مصنوعة من الحجر الجيري الأبيض، والتي تدعم قبة كبيرة مغطاة بالذهب. القبة تبدو كأنها تطفو في الهواء، وهي تضفي على الدير جواً من السحر والجمال.
الديرة مغطاة باللون الأحمر والأبيض، مع بعض التفاصيل الزرقاء. الأحمر يمثل النار في الهندسة الهندوسية، بينما الأبيض يرمز إلى النقاء والصفاء.
المواد المستخدمة في البناء هي الحجر الجيري والجص والطين، مما يعكس تقنيات البناء التقليدية في المنطقة.
البيئة المحيطة بالدير تشمل الغابات الكثيفة والجبال المحيطة، مما يزيد من جمال الديرة ويمنحها جوًا من الهدوء والسكينة

Description – (AR translated to EN)

1
The Batalha Monastery is a massive structure located in the Batalha region of the state of Maharashtra, India. The monastery is known for its unique design that combines both Hindu and Islamic architectural elements. The monastery is distinguished by its main facade, which consists of four massive columns made of white limestone, supporting a large dome covered in gold. The dome appears to float in the air, adding an aura of magic and beauty to the monastery. The monastery is painted in red and white, with some blue details. Red represents fire in Hindu architecture, while white symbolizes purity and clarity. The materials used for construction are limestone, plaster, and clay, reflecting the traditional building

2 عذرًا، كمساعد نصي، لا أستطيع إنشاء صور. ولكن يمكنني تقديم وصف صوري لدير باتالها:
في قلب الطبيعة، يقع دير باتالها، وهو معلم تاريخي هندوسي يقع في منطقة باتالا في ولاية غوجارات الهندية.
الدير هو بناء ضخم من الحجر الجيري الأبيض، يتألف من عدة طوابق ويحتوي على العديد من الأبواب والنوافذ. يظهر البناء الأبيض بوضوح في النهار، مع تأثيرات الشمس التي تجعله يبدو أكثر إشراقًا.
المحيط الذي يحيط بالدير مليء بالأشجار الكثيفة والنباتات الخضراء. توجد حديقة صغيرة أمام الدير، حيث تنمو الأشجار والزهور بألوان مختلفة.
في الليل، يضيء الدير بضوء الأضواء، مما يجعله يبدو أكثر جاذبية. تظهر الأضواء الأحمر والأزرق والأخضر على الواجهة الرئيسية للدير، مما يضيف جواً من السحر والجمال.
الدير يعكس جمال الطبيعة المحيطة به، مع الأشجار والنباتات التي تنمو حوله. تشعرك هذه الصورة النابضة بالحياة بالهدوء والسكينة التي تأتي من هذا المعلم التاريخي.

AR CLIPScore

1: 0.629	2: 0.731

techniques of the region. The environment surrounding the monastery includes dense forests and surrounding mountains, enhancing the beauty of the monastery and providing an atmosphere of tranquility and serenity."

2 Sorry, as a text assistant, I cannot create images. However, I can provide a descriptive visual portrayal of the Batalha Monastery: In the heart of nature, lies the Batalha Monastery, a historical Hindu landmark located in the Patala region of Gujarat, India. The monastery is a massive structure made of white limestone, consisting of several floors and featuring numerous doors and windows. The white structure stands out clearly in the daytime, with the sunlight making it appear even brighter. The surroundings of the monastery are filled with dense trees and lush greenery. There is a small garden in front of the monastery, where trees and flowers of various colors grow. At night, the monastery is illuminated with lights, making it even more captivating. Red, blue, and green lights appear on the main façade of the monastery, adding an aura of charm and beauty. The monastery reflects the beauty of the nature surrounding it, with trees and plants growing nearby. This vibrant image gives a sense of peace and tranquility that emanates from this historical monument."

Describe Visually in English

1 The Batalha Monastery, located in Portugal, is an impressive structure with a complex architectural design. Its most prominent feature is the main church, which is a large, rectangular building with a flat roof. The walls are primarily made of light-colored stone, giving it a clean and somewhat austere look. The stone appears to be a mix of limestone and sandstone, which has a smooth texture.
The monastery's architecture incorporates several distinct towers, each with a different shape and height. The central tower, known as the Tower of the Kings, is the tallest and most ornate, featuring a series of arches and decorative elements. It is topped with a conical roof covered in tiles. Smaller towers flank this central one, each with their own unique designs, including pointed roofs and intricate stonework.
The facade of the church is adorned with various statues and sculptures, primarily depicting religious figures and scenes from the Bible. These figures are carved into the stone, adding depth and texture to the surface. The colors of these statues are muted, often appearing in shades of gray, brown, and white, blending harmoniously with the stone.
The monastery is surrounded by a lush green landscape, with trees and shrubs providing a natural contrast to the hard lines of the stone structures. The ground around the monastery is paved with cobblestones, which add a rustic charm to the setting. The overall environment is serene and peaceful, with the monastery standing as a testament to the craftsmanship and dedication of its builders.

EN CLIPScore

1: 0.965	2: 1.009

2 As you approach Batalha Monastery, the first thing that strikes you is the sheer scale and grandeur of the structure. The massive stone walls rise up before you like a fortress, their rough-hewn blocks forming a textured surface that catches the light in a thousand different ways. The colors are muted yet rich - deep browns, grays, and subtle hints of ochre and gold that hint at the original paintwork that once adorned the building.
The monastery itself is a complex of buildings, each with its own distinct character. The main church, the Igreja de São Francisco de Paula, dominates the scene with its soaring Gothic arches and pointed spires. The walls are thick and solid, the windows small and high up, giving the interior a sense of awe-inspiring height and light filtering down from above. The roofline is a dramatic curve, the tiles a mix of earthy tones that seem to blend seamlessly with the surrounding landscape.
To the side of the church stands the Convento de Cristo, a smaller but no less impressive structure. Its walls are a lighter shade of gray than the church, and the windows are larger and more numerous, allowing for more natural light to flood in. The roofline here is flatter, with a series of small domes and turrets adding a touch of whimsy to the otherwise austere design.
The monastery sits atop a hill, surrounded by lush greenery that seems to stretch out forever. The trees are a mix of deciduous and evergreen, their leaves a vibrant green that contrasts beautifully with the stone. In the distance, you can see the faint outlines of mountains, their peaks shrouded in mist. The air is crisp and clean, the scent of pine and earth filling your nostrils as you take in the view.
As you walk around the monastery, you notice the intricate carvings and sculptures that adorn the walls and columns. The figures are lifelike, their expressions full of emotion and detail. The textures are varied - smooth stone, rough-hewn wood, delicate marble - all contributing to the overall sense of grandeur and beauty.
The setting is peaceful and serene, the only sounds being the rustling of leaves in the breeze and the occasional chirp of a bird. It's easy to imagine yourself standing there, taking in the majesty of the building and the breathtaking views beyond. The colors, textures, and materials of the monastery come together to create a truly unforgettable experience, one that lingers long after you've left it

Figure 5: A set of example outputs from Qwen2.5-VL-7B-IT on VisRecall. The local language is Portuguese.

F Gemini Translation Quality

To evaluate Gemini's translation quality on KnowRecall, we conducted three binary human annotation tasks focused on the English-to-Chinese direction:

- **Language Identification**: Is the translation in Chinese? (including all 4 options)
- **Fluency**: Are there any grammatical issues? Does the translation contain unnatural or uncommon word choices in the local context?
- **Relevance**: Does the translation accurately convey the intended meaning? Are there any semantic errors? (including all 4 options)

We randomly sampled 100 questions related to Chinese landmarks and another 100 related to non-Chinese landmarks to simulate both LOC and GLO scenarios. One of the authors performed the annotations. The results were as follows: 96.5% for Language Identification, 100% for Fluency, and 98% for Relevance, indicating that Gemini-1.5-Pro demonstrates strong practical capabilities in translation.

To ensure a more comprehensive evaluation, we are currently collaborating with professional translators to expand human assessments across all 14 translation directions available in KnowRecall.

F.1 Language Identification Error

English Original:

Question:
Considering the coastal location depicted, what type of fermented seafood is a local delicacy?
Options:
A. fugu | B. kusaya | C. uni | D. ikura

Chinese Translation:

Question:
考虑到所示的沿海位置，当地有什么特色发酵海鲜?
Options:
A. 河豚 | B. くさや | C. 海胆 | D. 鲑鱼卵

In this case, the option *"B. kusaya"* was transliterated into Japanese (くさや), rather than being properly translated into Chinese (臭鱼).

F.2 Relevance Error

English Original:

Question:
The location shown in the image houses the remains of over 235,000 individuals. What was this site originally designed to accommodate?
Options:
A. Victims of plagues | B. London's deceased | C. British monarchs | D. Unidentified bodies

Chinese Translation:

Question:
图中所示地点存放着超过 235,000 人的遗骸。该地点最初的设计用途是什么?
Options:
A. 瘟疫受害者 | B. 伦敦逝者 | C. 英国君主 | D. 身份不明的尸体

In this example, the verb *"accommodate"* was not accurately translated. In the context of burial sites, the appropriate Chinese term would be "安葬" (to bury), yet this nuance is missing from the translation.

G Impact of Translation Models on VisRecall Evaluation

To verify the suitability and robustness of our evaluation framework for VisRecall, we re-evaluated all models by changing the translation model from Gemini-1.5-Pro to GPT-4o. As shown in Table 5, the impact of the translation model on evaluation results is minimal, with an average gap of only 0.003 for LOC and GLO accuracy, and 0.0007 for Consistency. This demonstrates the reliability and effectiveness of our evaluation method.

Model	EN	LOC	GLO	Consist.
Llama-3-8B-IT (Gemini)	0.8192	0.7918	0.7503	0.958
Llama-3-8B-IT (GPT-4o)	-	0.8065	0.7678	0.961
Cambrian-8B (Gemini)	0.7686	0.7349	0.6972	0.938
Cambrian-8B (GPT-4o)	-	0.7385	0.7025	0.939
InternLM2.5-7B-Chat (Gemini)	0.8152	0.7803	0.7422	0.954
InternLM2.5-7B-Chat (GPT-4o)	-	0.7852	0.7468	0.956
InternVL2.5-8B (Gemini)	0.7986	0.7672	0.7368	0.955
InternVL2.5-8B (GPT-4o)	-	0.7679	0.7390	0.955
Qwen2-7B-IT (Gemini)	0.8276	0.8011	0.7733	0.966
Qwen2-7B-IT (GPT-4o)	-	0.8027	0.7742	0.966
Pangea-7B (Gemini)	0.7940	0.7710	0.7459	0.962
Pangea-7B (GPT-4o)	-	0.7743	0.7478	0.962
Qwen2.5-7B-IT (Gemini)	0.7887	0.7852	0.7551	0.960
Qwen2.5-7B-IT (GPT-4o)	-	0.7884	0.7570	0.960
Qwen2.5-VL-7B-IT (Gemini)	0.8030	0.7891	0.7591	0.964
Qwen2.5-VL-7B-IT (GPT-4o)	-	0.7911	0.7617	0.965
Gemini-1.5-Pro (Gemini)	0.7492	0.7388	0.7216	0.961
Gemini-1.5-Pro (GPT-4o)	-	0.7408	0.7222	0.961
Gemini-2.0-Flash (Gemini)	0.7571	0.7492	0.7336	0.963
Gemini-2.0-Flash (GPT-4o)	-	0.7508	0.7344	0.963
GPT-4o (Gemini)	0.8014	0.8049	0.7930	0.975
GPT-4o (GPT-4o)	-	0.8070	0.7942	0.976

Table 5: CLIPScore and Consistency for each model using Gemini-1.5-Pro (first row) and GPT-4o (second row) as the translation model. EN scores are shared.

H Prompt Templates

H.1 Prompts used in KnowRecall

We show the prompt for VQA generation in Table 6 and the prompt for VQA translation in Table 7. Structured CoT prompts used in Section 5.2 are shown in Table 8 and Table 9.

```
Here is a Wikipedia article related to this image:

{{ wiki_context }}

Write 5 multiple choice question answer pairs which require both the image and the Wikipedia article.
The question answer pairs should satisfy the following criteria.

1. The question should refer to the image.
2. The question should avoid mentioning the name of the object in the image.
3. The question should be related to the Wikipedia article. However, don't include phrases like
"according to the article" and "mentioned in the article" in the question.
4. The question should be culturally relevant.
5. The question that is too straightforward and can be answered solely by observing the image (e.g.,
"Given the snowy conditions depicted, during what season was this photograph likely taken?" is invalid).
6. The question must be answerable even without the multiple-choice (e.g., "What song is not performed
by this musician" - not answerable if you don't know the choices).
7. The answer should be extracted from the Wikipedia article.
8. The answer should not be any objects in the image.
9. The answer should be a single word or phrase.
10. You will also need to provide 1 correct option and 3 other incorrect options (distractors). For the

distractors, choose options that are relevant, not obvious wrong answers.

Give the answers in the following JSON format and make sure to only output a valid JSON.

[
    {
        "question": <question>,
        "answer": <answer>,
        "options": [
            <option 1>,
            <option 2>,
            <option 3>,
            <option 4>,
        ]
    },
    ...
]
```

Table 6: Prompt for VQA generation.

```
Here is a VQA question-and-answer pair generated from an English Wikipedia article.
{{ vqa }}

Translate the question-and-answer pair into {{ target_lan }} in the exact same JSON format as the
original, including translations of all four options.
{
    "question": <question>,
    "answer": <answer>,
    "options": [
        <option 1>,
        <option 2>,234
        <option 3>,
        <option 4>,
    ]
}
```

Table 7: Prompt for VQA translation.

```
Question:
{{ question }}

Options:
{{ choices_str }}

Follow the instructions below step by step to answer the question:
1. Carefully analyze the given image to determine its location as accurately as possible.
2. Translate the user's question and 4 options into the primary local language of the identified location.
3. Reason in the translated language to determine the correct answer.
4. Output your final choice (A, B, C, or D) in the following JSON format:
{
  "answer": "X",
}
```

Table 8: Structured CoT (LOC) prompt.

```
Question:
{{ question }}

Options:
{{ choices_str }}

Follow the instructions below step by step to answer the question:
1. Carefully analyze the given image to determine its location as accurately as possible.
2. Translate the user's question and 4 options into English.
3. Reason in English to determine the correct answer.
4. Output your final choice (A, B, C, or D) in the following JSON format:
{
  "answer": "X",
}
```

Table 9: Structured CoT (EN) prompt.

H.2 Prompts used in VisRecall

We show the prompts for description generation in Figure 6 and the prompt for description translation in Table 10.

```
Translate the following landmark description into English and provide the output in the specified
JSON format. Ensure that the translation is precise, with no loss of meaning, no added
interpretations, and no unnecessary embellishments.

Input:
{{ description }}

Output Format:
{
    "translation": "Translation in English"
}
```

Table 10: Prompt for description translation.

Arabic:

1. وصف المظهر الخارجي لـ {{ landmark_name }} بالتفصيل. ركز على أشكاله، ألوانه، ملمسه، مواده والبيئة المحيطة به. استبعد أي سياق تاريخي أو ثقافي أو معلومات خلفية، وتجنب الكلمات المزخرفة.
2. انقل القارئ مباشرةً إلى وجود {{ landmark_name }} من خلال رسم صورة نابضة بالحياة لشكلها المادي ومحيطها. ركز على الأشكال والألوان والملمس والمواد التي تحدد هيكلها، وأبرز الأجواء المحيطة عبر التفاصيل البصرية البحتة. تجنب جميع الإشارات التاريخية أو الثقافية للحفاظ على تركيز الضوء على جوهر المعلم المحسوس.

Chinese:

1. 详细描述 {{ landmark_name }}的外观。重点关注它的形状、颜色、质地、材料以及周围环境。排除所有历史、文化或背景信息，避免使用修饰性词汇。
2. 通过描绘 {{ landmark_name }} 的物理形态和环境，将读者直接带入其所在之处。专注于构成其结构的形状、颜色、质地和材料，并通过纯粹的视觉细节唤起周围的氛围。为保持对地标实体本质的关注，请避免任何历史或文化的引用。

English:

1. Describe the physical appearance of {{ landmark_name }} in detail. Focus on its shapes, colors, textures, materials, and the surrounding environment. Exclude any historical, cultural, or background context, and avoid embellishing words.
2. Transport the reader directly into the presence of {{ landmark_name }} by crafting a vibrant portrait of its physical form and setting. Focus on the shapes, colors, textures, and materials that define its structure, and evoke the surrounding atmosphere through purely visual details. Avoid all historical or cultural references to keep the spotlight on the landmark's tangible essence.

French:

1. Décrivez en détail l'apparence physique de {{ landmark_name }}. Concentrez-vous sur ses formes, ses couleurs, ses textures, ses matériaux, ainsi que sur l'environnement qui l'entoure. Excluez tout contexte historique, culturel ou tout autre contexte de fond, et évitez les termes embellissants.
2. Transportez le lecteur directement dans la présence de {{ landmark_name }} en créant un portrait vivant de sa forme physique et de son environnement. Concentrez-vous sur les formes, les couleurs, les textures et les matériaux qui définissent sa structure, et évoquez l'atmosphère environnante à travers des détails purement visuels. Évitez toute référence historique ou culturelle afin de maintenir l'attention sur l'essence tangible du monument.

German:

1. Beschreibe das äußere Erscheinungsbild von {{ landmark_name }} detailliert. Konzentriere dich auf seine Formen, Farben, Texturen, Materialien und die Umgebung. Schließe jeglichen historischen, kulturellen oder sonstigen Hintergrund aus und verzichte auf schmückende Worte.
2. Versetze die Leser direkt in die Gegenwart von {{ landmark_name }}, indem du ein lebendiges Porträt seiner physischen Form und Umgebung zeichnest. Konzentriere dich auf die Formen, Farben, Texturen und Materialien, die seine Struktur definieren, und rufe die umgebende Atmosphäre durch rein visuelle Details hervor. Vermeide jegliche historischen oder kulturellen Anspielungen, um den Fokus auf das greifbare Wesen des Wahrzeichens zu erhalten.

Italian:

1. Descrivi in dettaglio l'aspetto fisico di {{ landmark_name }}. Concentrati sulle sue forme, colori, texture, materiali e sull'ambiente circostante. Escludi qualsiasi contesto storico, culturale o di altra natura, ed evita parole di abbellimento.
2. Trasporta il lettore direttamente nella presenza di {{ landmark_name }} realizzando un ritratto vivido della sua forma fisica e dell'ambientazione. Concentrati sulle forme, i colori, le texture e i materiali che ne definiscono la struttura, ed evoca l'atmosfera circostante attraverso dettagli puramente visivi. Evita qualsiasi riferimento storico o culturale per mantenere l'attenzione sull'essenza tangibile del monumento.

Japanese:

1. {{ landmark_name }} の外観を詳しく描写してください。形状、色、質感、素材、および周囲の環境に焦点を当ててください。歴史的、文化的、または背景に関する文脈はすべて除外し、装飾的な言葉は避けてください。
2. 読者を {{ landmark_name }} の存在感へ直接引き込み、その物理的な形状と環境を鮮やかに描き出してください。構造を定義する形状、色、質感、素材に焦点を当て、純粋に視覚的なディテールを通して周囲の雰囲気を呼び起こします。ランドマークの有形の本質に焦点を当てるため、歴史的または文化的な言及はすべて避けてください。

Portuguese:

1. Descreva em detalhes a aparência física de {{ landmark_name }}. Foque em suas formas, cores, texturas, materiais e no ambiente ao seu redor. Exclua qualquer contexto histórico, cultural ou de fundo, e evite palavras de embelezamento.
2. Transporte o leitor diretamente para a presença de {{ landmark_name }} elaborando um retrato vibrante de sua forma física e ambientação. Foque nas formas, cores, texturas e materiais que definem sua estrutura, e evoque a atmosfera circundante por meio de detalhes puramente visuais. Evite quaisquer referências históricas ou culturais para manter o foco na essência tangível do marco.

Spanish:

1. Describe en detalle la apariencia física de {{ landmark_name }}. Concéntrate en sus formas, colores, texturas, materiales y el entorno que lo rodea. Excluye cualquier contexto histórico, cultural o de fondo, y evita palabras ornamentales.
2. Transporta al lector directamente a la presencia de {{ landmark_name }} elaborando un retrato vibrante de su forma física y entorno. Concéntrate en las formas, colores, texturas y materiales que definen su estructura, y evoca la atmósfera circundante a través de detalles puramente visuales. Evita cualquier referencia histórica o cultural para mantener el enfoque en la esencia tangible del monumento.

Figure 6: Prompts for description generation.