VLR-Bench: Multilingual Benchmark Dataset for Vision-Language Retrieval Augmented Generation

Hyeonseok Lim*, Dongjae Shin‡*, Seohyun Song, Inho Won‡
Minjun Kim, Junghun Yuk, Haneol Jang**†, KyungTae Lim†**
Seoul National University of Science and Technology (SeoulTech)
‡SeoulTech & Teddysum
**Hanbat National University
{gustjrantk,dylan1998,alexalex225225,wih1226,mjkmain}@seoultech.ac.kr,
20191780@hanbat.ac.kr, hejang@hanbat.ac.kr, ktlim@seoultech.ac.kr

Abstract

We propose the VLR-BENCH, a visual question answering (VQA) benchmark for evaluating vision language models (VLMs) based on retrieval augmented generation (RAG). Unlike existing evaluation datasets for external knowledge-based VQA, the proposed VLR-BENCH includes five input passages. This allows testing of the ability to determine which passage is useful for answering a given query, a capability lacking in previous research. In this context, we constructed a dataset of 32,000 automatically generated instruction-following examples, which we denote as VLR-IF. This dataset is specifically designed to enhance the RAG capabilities of VLMs by enabling them to learn how to generate appropriate answers based on input passages. We evaluated the validity of the proposed benchmark and training data and verified its performance using the state-of-the-art Llama3-based VLM, the Llava-Llama-3 model. The proposed VLR-BENCH[1] and VLR-IF[2] datasets are publicly available online.

1 Introduction

The search for external knowledge is very important for VLMs because it is often impossible to find answers directly from images in response to user queries (Marino et al., 2019). Previous studies attempted to incorporate external knowledge into VLMs. Among these efforts, dense passage retrieval (Karpukhin et al., 2020) has been used to search for documents related to queries in an attempt to solve this problem (Luo et al., 2021; Gao et al., 2022). However, as Lin and Byrne (2022) pointed out, these models face challenges in determining whether the retrieved documents are useful for answering queries.Following this, the proposed RA-VQA (Lin and Byrne, 2022) introduced an approach that simultaneously conducts searches and question-answering to overcome these drawbacks. However, since the study primarily focused on the RAG configuration, evaluating how the VLM utilized the search results remained challenging.

To address these issues, we propose a Vision Language-RAG Benchmark (VLR-BENCH) and training data to evaluate the Retrieval-Augmented Generation (RAG) capabilities of VLMs (Lewis et al., 2021). VLR-BENCH consists of 300 datasets composed of problems that are difficult to solve without external knowledge. The data were structured as an image-query-passage-output, and unlike conventional VQA datasets, each dataset contained five distinct passages. Only two passages contained direct information that could resolve the queries. This allows us to test the ability, which has been lacking in previous research, to determine which passages are useful for answering queries.

In this study, we developed the VLR Instruction Following (VLR-IF) training data for VLM RAG based on the data generation method proposed by LLaVA (Liu et al., 2024) and assessed its utility. We validated the proposed VLR-BENCH and VLR-IF training data based on the following three research questions: (1) Does the proposed VLR-BENCH require external knowledge retrieval to be solved? (2) How does the proposed training data impact external knowledge utilization? (3) How effectively can public VLMs and commercial models resolve queries that require retrieval?

In this study, we conducted a baseline performance evaluation of VLR-BENCH using the most recently released vision language models in the LLAVA-LLAMA-3 series (Contributors, 2023) and GPT-4o (OpenAI et al., 2024). The contributions of this study can be summarized as follows:

- We propose multilingual RAG evaluation data, VLR-BENCH, and training data, VLR-IF, for

*These authors contributed equally.
†Corresponding authors.
[1]https://huggingface.co/datasets/MLP-KTLim/VLR-Bench
[2]https://huggingface.co/datasets/MLP-KTLim/VLR-IF

Proceedings of the 31st International Conference on Computational Linguistics, pages 6150–6168
January 19–24, 2025. ©2025 Association for Computational Linguistics

the VLMs.
- Through in-depth analysis, we prove the actual effect of our dataset.

2 Related Work

VLM Benchmark Datasets. In the VLM benchmark, OK-VQA (Marino et al., 2019) is a key open-domain VQA dataset that uses external knowledge from Wikipedia. Subsequently, A-OKVQA (Schwenk et al., 2022) and S3VQA (Jain et al., 2021), which included justifications for answers, were derived from OK-VQA. Additionally, datasets targeting specific domains have appeared; for instance, K-VQA (Shah et al., 2019), which intensively utilizes personal information, and ViQuAE (Lerner et al., 2022), which uses object information, were proposed as evaluation datasets. Furthermore, VQA models utilizing knowledge graphs have been proposed, notably GQA (Hudson and Manning, 2019), which uses scene graph knowledge and its multilingual expansion (xGQA (Pfeiffer et al., 2022) and BOK-VQA (Kim et al., 2024)). In a different context, datasets providing passages for evaluating the RAG capabilities of VLMs have recently emerged. Notable examples include InfoSeek (Chen et al., 2023) and Encyclopedic VQA (Mensink et al., 2023). These datasets provide passages or entire documents, resulting in performance variations based on the document retrieval ability. Detailed information on these external knowledge-based VLM benchmark datasets, as well as their differences from the proposed VLR-BENCH, can be found in Appendix D.3.

3 Proposed RAG Dataset for VLMs

Benchmarks related to the use of external knowledge by VLMs, as discussed in Section 2, particularly InfoSeek and Encyclopedic VQA, typically provide single gold-standard evidence to resolve queries. However, real-world RAG-based systems generate answers by incorporating multiple retrieved results (e.g., Top-5). A significant challenge arises when plausible but incorrect information is retrieved as external knowledge. Therefore, when VLM models use RAG, it is essential to evaluate (1) how accurately external knowledge is retrieved and (2) the model's ability to generate correct answers despite the existence of incorrect information. In this context, we propose VLR-BENCH, which simultaneously considers the correct selection of external knowledge and answers-generated by VLMs.

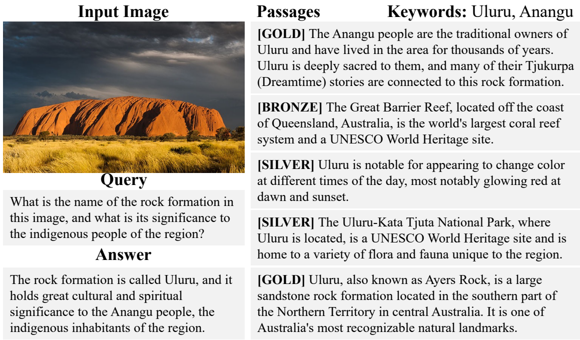


Figure 1: An example of VLR-BENCH data sample.

In addition, we introduce a construction method for the VLR-IF dataset designed to enhance the ability of VLMs to select external knowledge.

3.1 VLR-Bench Dataset

VLR-BENCH was constructed to evaluate whether VLMs can use the correct external knowledge to generate accurate responses to query. We constructed a parallel corpus of 300 datasets: 150 based on general knowledge and 150 based on cultural data from English, Chinese, and Korean. Detailed examples of the data are provided in Appendix A.

Image Selection. Images are crucial within this dataset. The diversity of categories among the selected images is essential for depicting a range of external knowledge. Considering these factors, we manually curated 150 images from BOK-VQA, developed explicitly for open-world QA purposes.

We manually extracted 150 images from the 10 categories proposed by BOK-VQA, with 15 images each from the object-centric, atmosphere-centric, and relation-centric categories. In addition, We collected 150 images of different languages' cultural backgrounds from Wikimedia Commons under the same conditions as BOK-VQA.

Question Selection. The question selection process used GPT-4o to receive recommendations for high-quality question-answer pairs. We input images into GPT-4o and requested them to generate ten queries, two essential pieces of external knowledge required to resolve these queries, and descriptive answers. To ensure the validity of the model verification, we imposed the following conditions: (1) The generated data should consist of question-answer pairs that cannot be resolved with the image alone. (2) Image information should not be explicitly evident in the questions to ensure that queries cannot be resolved using external knowl-

Lang.	Model	VLR-IF			With Passages					Without Passages				
		EN	ZH	KO	KMS	R-2	R-L	BLEU	B-Score	KMS	R-2	R-L	BLEU	B-Score
EN	LLAVA1.5 (Liu et al., 2024)	✗	✗	✗	**88.4**	26.0	37.2	14.7	76.3	25.6	17.8	30.4	9.1	73.4
	LLAVA-LLAMA-3	✗	✗	✗	79.2	25.4	38.8	13.5	79.3	20.4	12.2	23.8	6.1	73.4
	LLAVA-LLAMA-3+VLR-IF(EN)	✓	✗	✗	85.6	30.1	46.4	20.9	81.5	20.4	19.1	29.9	8.6	69.1
	X-LLAVA (Shin et al., 2024)	✗	✗	✗	80.4	28.1	42.2	16.9	80.1	20.8	17.5	31.9	9.6	74.3
	X-LLAVA+VLR-IF(EN)	✓	✗	✗	82.4	29.4	44.2	20.1	80.7	20.4	19.7	35.4	12.7	77.5
	X-LLAVA+VLR-IF(EN+KO)	✓	✗	✓	83.2	30.2	45.2	20.6	81.0	18.4	20.4	36.3	14.5	77.1
	QWEN-VL-CHAT	✗	✗	✗	84.8	32.8	47.4	20.6	82.1	31.2	20.0	34.1	9.7	77.5
	GPT-4o (OpenAI et al., 2024)	✗	✗	✗	85.6	**42.6**	**57.9**	**32.8**	**85.6**	**61.6**	**35.6**	**52.1**	**26.2**	**83.7**
ZH	QWEN-VL-CHAT (Bai et al., 2023)	✗	✗	✗	75.6	51.6	56.3	33.8	84.0	10.8	28.9	37.2	18.2	75.4
	QWEN-VL-CHAT+VLR-IF(ZH)	✗	✓	✗	72.4	**59.0**	**63.4**	**42.9**	**86.2**	16.0	30.4	37.8	18.1	77.4
	GPT-4o	✗	✗	✗	**80.4**	56.9	62.3	41.6	**86.2**	**36.0**	**36.6**	**42.9**	**24.6**	**80.3**
KO	X-LLAVA	✗	✗	✗	59.6	27.0	35.2	15.2	78.4	6.0	18.0	28.0	8.6	74.2
	X-LLAVA+VLR-IF(KO)	✗	✗	✓	63.6	35.7	44.1	24.9	81.0	6.8	22.4	32.9	14.9	77.0
	X-LLAVA+VLR-IF(EN+KO)	✓	✗	✓	62.4	36.0	44.6	24.2	81.7	0.8	4.7	15.2	5.14	64.5
	GPT-4o	✗	✗	✗	**83.6**	**51.9**	**55.2**	**37.2**	**84.4**	**31.6**	**35.9**	**39.0**	**24.9**	**79.7**

Table 1: Overall experiment results on VLR-BENCH depending on its language. (R: Rouge and B-Score: Bert-Score)

edge alone. The data produced consisted of queries related to each sample image, two pieces of external knowledge necessary to solve the queries, and a descriptive answer. At this stage, we selected the most suitable samples from the ten recommended query-knowledge pairs and conducted a preliminary review to verify that all the data consisted of queries requiring external knowledge.

Generation of Additional External Knowledge VLR-BENCH consists of five pieces of external knowledge. Among these, two are directly referenced when generating answers for the actual images and questions, referred to as '`Gold Passage`', which were already reviewed in the previous stage. Two of the five passages relate to the theme of the image or question but diverge from the central theme of the answer, termed '`Silver Passage`'. The last one, unrelated to the image and the question, is designated as '`Bronze Passage`'. At this stage, we generated two silver passages and one bronze passage. Three annotators directly reviewed the data derived through this process for the question-answer pairs, external knowledge, and descriptive answers. Specifically, errors in the generated external knowledge or knowledge with unclear sources were replaced with new information by annotators (see Appendix A.2). Finally, each annotator extracted the two essential keywords necessary to resolve the questions. Each sample comprises five elements: an image, a query, five pieces of knowledge, a descriptive answer, and two keywords. Examples of the data are shown in Figure 1.

3.2 VLR-IF Dataset

To address the proposed benchmark, we designed instruction-following data to enhance the utilization of external knowledge using VLMs. As previously proposed, we generated data using the same GPT-based method for question-external knowledge-answer creation. Initially, we randomly selected 9K COCO (Lin et al., 2015) images and generated a 'valid passage' related to each image. Subsequently, we randomly extracted external knowledge from different data samples for use as 'invalid passages', thus contrastively constructing datasets using a combination of valid and invalid passages. The VLR-IF dataset was constructed in parallel for three languages: English, Chinese, and Korean, with each language comprising 32K data samples. The specific process for constructing the datasets is described in Appendix B.2.

LMM	LLM	#PT	#VIT	Language
LLAVA1.5	Llama2-13 B	558 K	665 K	En,Ko,Zh
LLAVA-LLAMA-3	Llama3-8 B	1.2 M	1.2 M	En
X-LLAVA	Llama2-13 B	1.2 M	407 K	En,Ko
QWEN-VL	Qwen 7 B	1.4 B	350 K	Zh, En

Table 2: VLMs for evaluation on VLR-BENCH

4 Experiments and Analysis

We selected the top-performing models for each language for our experiments. Table 2 presents the base models, pre-training volumes, and visual instruction tuning (VIT) training volumes for the models used in this experiment. The VLR-BENCH task involves generating long-form answers to the given queries. As described in Section 3, two keywords were manually annotated for each query. Therefore, these keywords in the model-generated long-form answers allow for some degree of quantitative evaluation, defined as the keyword-matching score (KMS). We considered a response correct only when both answer keywords were accurately identified. However, because the KMS performance may improve as the generated response lengthens, it is used as a reference indicator rather than an exact performance measure. To compensate for this, a comprehensive evaluation should be conducted using metrics that account for sentence length, such as Rouge (Lin, 2004), BLEU (Papineni et al., 2002), and BERT-Score (Zhang* et al., 2020).

Table 1 presents the evaluation results for each language-specific model. If the proposed VLR-IF data were used for training the models, it was denoted as +VLR-IF; the hyperparameters used in this case can be found in Appendix C.

4.1 Experiment Results

Diversity in Performance Evaluation. Upon examining the English KMS performance in the With Passage section of Table 1, it can be observed that the performance of LLAVA1.5 closely mirrors that of GPT-4o. This raises the question of whether LLAVA1.5 truly makes accurate predictions. The answer is no. The task involves generating long-form answers, and LLAVA1.5 often directly outputs the received external information, resulting in lengthy responses. Although such responses achieve high KMS performance, they also contain external knowledge irrelevant to the query, leading to lower BLEU and Rouge scores.

The Impact of Use of External Knowledge. VLR-BENCH allows for evaluations in scenarios where external knowledge is provided, as each problem is accompanied by five pieces of external knowledge. Table 1 presents the VLR-BENCH evaluation results based on the availability of external knowledge for each model. Notably, the performance of the X-LLAVA model dropped by an average of 37.72% for R-2 in English compared to when external knowledge was provided. These results suggest that the VLR-BENCH dataset contains queries that require external knowledge.

The Impact of VLR-IF Training. We conducted experiments to assess the utility of the VLR-IF data using the baseline LLAVA-LLAMA-3 and its version enhanced by VLR-IF training. According to the results in Table 1, the model trained with the VLR-IF data showed a 22.67% performance improvement over the baseline model when external knowledge was provided. This significant enhancement suggests that the VLR-IF training data effectively boosts the ability to select and utilize external knowledge. Finally, we examined whether VLR-IF could positively impact other evaluation datasets, using the InfoSeek (Chen et al., 2023) benchmark as a reference. The results indicated a 3.6% performance improvement with the application of VLR-IF (see Appendix C.3).

Lang.	Passage-type	BERT Score f1	Rouge-1	Rouge-2	Rouge-L
Passages & Ground-truth output					
EN	Gold	**78.04**	**44.96**	**20.98**	**31.92**
	Silver	72.11	25.59	5.978	18.92
	Bronze	67.81	22.10	2.299	17.04
Passages & Questions					
EN	Gold	**66.55**	**28.42**	**4.67**	**18.93**
	Silver	65.70	21.48	1.97	15.79
	Bronze	64.48	24.11	2.74	17.26

Table 3: Correlation analysis between Passages, Ground Truth, and Questions for the English cases.

Comparing GPT-4o with Open Models. We conducted experiments to test if GPT-4o could solve VLR-BENCH problems without external knowledge. The results from the "Without Passages" section in Table 1 show that GPT-4o outperformed QWEN-VL-CHAT by an average of 17.33 points without external knowledge. However, with passages provided, the performance gap narrowed to an average of 7.36 points. This indicates that VLR-BENCH is a challenging benchmark without external knowledge, and open-source models can improve with passage-retrieval capabilities.

4.2 Analysis

In this section, we present an in-depth analysis to determine whether the VLR-BENCH is a suitable dataset for evaluating model's ability to utilize information. To this end, we measured the BERT-score and Rouge scores (R-1, R-2, R-L) between passage types, questions, and ground-truth outputs. The results presented in Table 3 show that the ground truth output correlates most strongly with the `Gold - Silver - Bronze Passage` in descending order. This trend substantiates the effective use of gold passages in deducing answers to the VLR-BENCH, indicating that the appropriate utilization of externally sourced knowledge through images is crucial for answering queries. On the other hand, an examination of the passages and question results reveals no clear trend, as Bronze's Rouge-1 score is higher than Silver's, suggesting that selecting suitable external knowledge based solely on the query can be challenging. This implies that understanding the images is necessary.

5 Conclusion

In this study, we propose VLR-BENCH for evaluating RAG-based VLMs, and VLR-IF for performance enhancement. The proposed benchmark differs from existing external knowledge-based VLM evaluation datasets in the following ways. (1) It consists of problems that are difficult to solve without external knowledge. (2) It includes five different passages, allowing the test of an ability not covered in previous research to determine which passages are useful for answering queries. The training data were designed as multilingual evaluation data that could simultaneously assess English, Chinese, and Korean, enhancing their utility.

6 Limitations

In this study, we proposed a benchmark and corresponding training data to evaluate the RAG capabilities of VLMs. The benchmark allows for the evaluation of both retrieval and generation abilities. However, there are still two issues that remain:

Absence of Image Search Capability. Ultimately, the ability to perform image searches is crucial for accurately assessing the performance of the VLR-Bench. As mentioned in Table 1, the superior performance of GPT-4o over other public language models originates from the presence or absence of image search capabilities. Unfortunately, this study did not consider methods related to image search.

Lack of Diversity in Responses Due to Training Data Construction Costs. The method proposed in this study enabled the construction of training data at a very low cost. However, applying the same method to other languages still incurs costs, particularly when building test data, which can be expensive. Due to these cost constraints, annotation was performed by a single individual. While there could be multiple correct answers to the short-answer core keywords, due to budget limitations, responses were collected from only one person. Nevertheless, the final test data underwent a secondary review process to ensure data quality.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant, funded by the Korea government (MSIT) (No.RS-2024-00456709, A Development of Self-Evolving Deepfake Detection Technology to Prevent the Socially Malicious Use of Generative AI) and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea)& Gwangju Metropolitan City awarded to KyungTae Lim.

References

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

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *Preprint*, arXiv:2302.11713.

XTuner Contributors. 2023. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5057–5067. IEEE.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21. ACM.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

MinJun Kim, SeungWoo Song, YouHan Lee, Haneol Jang, and KyungTae Lim. 2024. Bok-vqa: Bilingual outside knowledge-based visual question answering via graph representation pretraining. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18381–18389.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno, and Jesús Lovón Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'22, New York, NY, USA. Association for Computing Machinery.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431.

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

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, Andre Araujo, and Vittorio Ferrari. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *ICCV*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin 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.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Dongjae Shin, Hyeonseok Lim, Inho Won, Changsu Choi, Minjun Kim, Seungwoo Song, Hangyeol Yoo, Sangmin Kim, and Kyungtae Lim. 2024. X-llava: Optimizing bilingual large vision-language alignment. *Preprint*, arXiv:2403.11399.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Appendix

A VLR-Bench

A.1 VLR-Bench Examples

Overall The following figures are examples from VLR-BENCH. Each example consists of a question, an answer, keywords, and passages. The "gold passage", which contains the information necessary to answer the question, is highlighted in yellow.

Q (EN) : What is the name of the monument in this image, and which president's face was carved first?
A (EN) : The monument is Mount Rushmore, and George Washington's face was carved first.
Keywords (EN) : Mount Rushmore, George Washington

Q (ZH) : 这张图片中的纪念碑叫什么名字，哪位总统的脸是第一个雕刻的？
A (ZH) : 这座纪念碑叫拉什莫尔山，第一个雕刻的总统脸是乔治·华盛顿。
Keywords (ZH) : 拉什莫尔山, 乔治·华盛顿

Q (KO) : 이 이미지에 있는 기념물의 이름은 무엇이며, 어느 대통령의 얼굴이 가장 먼저 새겨졌나요?
A (KO) : 이 기념물은 러시모어 산이며, 조지 워싱턴의 얼굴이 가장 먼저 새겨졌습니다.
Keywords (KO) : 러시모어 산, 조지 워싱턴

Passages

1) The sculptor behind Mount Rushmore was Gutzon Borglum, and the project involved over 400 workers.
2) The other presidents depicted on Mount Rushmore are Thomas Jefferson, Theodore Roosevelt, and Abraham Lincoln.
3) Mount Rushmore National Memorial is a massive sculpture carved into the granite face of Mount Rushmore in the Black Hills of South Dakota. It features the 60-foot heads of four U.S. presidents.
4) The faces of Mount Rushmore were carved between 1927 and 1941. George Washington's face was the first to be carved, starting in 1927.
5) The Eiffel Tower in Paris, France, was completed in 1889 and was the tallest man-made structure in the world until the completion of the Chrysler Building in New York City in 1930.

1) 拉什莫尔山上描绘的其他总统是托马斯·杰斐逊、西奥多·罗斯福和亚伯拉罕·林肯。
2) 拉什莫尔山国家纪念碑是一座雕刻在南达科他州黑山拉什莫尔山花岗岩上的巨大雕塑。它展示了四位美国总统的60英尺高的头像。
3) 法国巴黎的埃菲尔铁塔于1889年完工，直到1930年纽约市的克莱斯勒大厦完工之前，它一直是世界上最高的人工结构。
4) 拉什莫尔山的面孔是在1927年至1941年间雕刻的。乔治·华盛顿的面孔是第一个被雕刻的，从1927年开始。
5) 拉什莫尔山背后的雕塑家是古森·博格伦，该项目涉及400多名工人。

1) 러시모어 산에 묘사된 다른 대통령들은 토머스 제퍼슨, 시어도어 루즈벨트, 그리고 아브라함 링컨입니다.
2) 프랑스 파리의 에펠탑은 1889년에 완공되었고, 1930년 뉴욕 시의 크라이슬러 빌딩이 완공될 때까지 세계에서 가장 높은 인공 구조물이었습니다.
3) 러시모어 산 국립 기념물은 사우스다코타 주 블랙힐스의 러시모어 산 화강암 면에 새겨진 거대한 조각상입니다. 이 조각상은 네 명의 미국 대통령의 60피트 크기 머리를 특징으로 합니다.
4) 러시모어 산 조각의 뒤에는 구츠온 보글럼이 있었고, 이 프로젝트에는 400명 이상의 노동자들이 참여했습니다.
5) 러시모어 산의 얼굴들은 1927년부터 1941년 사이에 조각되었습니다. 조지 워싱턴의 얼굴이 1927년에 처음으로 조각되었습니다.

Figure 2: Examples of the created VLR-Bench data. (English culture)

Q (EN) : What breed of cattle is shown in this image, and what are their distinctive characteristics that make them suitable for harsh environments?
A (EN) : The breed of cattle shown in this image is the Highland cattle. They are known for their long horns and long, wavy, woolly coats that help them withstand harsh weather conditions, particularly in the Scottish Highlands.
Keywords (EN) : Highland cattle, Woolly coat

Q (ZH) : 这张图片中的牛是什么品种，它们有哪些独特特征使其适应恶劣环境？
A (ZH) : 这张图片中的牛是高地牛。它们以长角和长而波浪状的毛皮著称，这些特征帮助它们在恶劣的天气条件下生存，尤其是在苏格兰高地。
Keywords (ZH) : 高地牛, 波浪状毛皮

Q (KO) : 이 이미지에 나오는 소의 품종은 무엇이며, 어떤 독특한 특징 덕분에 가혹한 환경에서도 견딜 수 있나요?
A (KO) : 이 이미지에 나오는 소의 품종은 하이랜드 소입니다. 이들은 긴 뿔과 길고 곱슬거리는 털로 유명하며, 이러한 털은 특히 스코틀랜드 하이랜드의 가혹한 날씨를 견디는 데 도움을 줍니다.
Keywords (KO) : 하이랜드 소, 곱슬거리는 털

Passages

1) The dense coat of Highland cattle provides protection against cold, wet, and windy weather, making them particularly resilient in the Scottish Highlands.
2) Highland cattle are known for their docile temperament and are often used in conservation grazing to maintain natural landscapes.
3) Highland cattle are a Scottish breed of rustic beef cattle. They have long horns and long, wavy, woolly coats that are well-suited to harsh weather conditions.
4) The Eiffel Tower, located in Paris, France, was completed in 1889 and is one of the most recognizable structures in the world.
5) The breed's meat is highly prized for its flavor and tenderness, contributing to its popularity in gourmet cuisine.

1) 这种品种的肉因其风味和嫩度而备受推崇，在美食中非常受欢迎。
2) 位于法国巴黎的埃菲尔铁塔于1889年完工，是世界上最具辨识度的建筑之一。
3) 高地牛以其温顺的性格而闻名，常用于保护性放牧以维护自然景观。
4) 高地牛是苏格兰的一种乡村牛品种。它们有长长的角和长长的波浪状毛绒外套，非常适合恶劣的天气条件。
5) 高地牛的浓密毛皮提供了对寒冷、潮湿和多风天气的保护，使它们在苏格兰高地特别坚韧。

1) 하이랜드 소는 스코틀랜드의 소 품종으로, 시골에서 자라는 소입니다. 이 소들은긴 뿔과 길고 곱슬거리는 털을 가지고 있어 혹독한 날씨에 잘 적응합니다.
2) 하이랜드 소는 온순한 성격으로 유명하며, 자연 경관을 유지하기 위해 보존 방목에 자주 사용됩니다.
3) 하이랜드 소의 곱슬거리는 털은 추위, 습기, 바람으로부터 보호해주어, 스코틀랜드 고지대에서 특히 강인하게 살아남습니다.
4) 이 품종의 고기는 맛과 부드러움으로 높이 평가받아, 미식 요리에서 인기가 많습니다.
5) 프랑스 파리에 위치한 에펠탑은 1889년에 완공되었으며, 세계에서 가장 잘 알려진 구조물 중 하나입니다.

Figure 3: Examples of the created VLR-Bench data. (commonsense knowledge)

Data incorporating language-specific cultural aspects. VLR-BENCH comprises 150 datasets for each language, incorporating language-specific cultural aspects. The benchmark is designed to include queries that require an understanding of the respective culture to accurately select the correct information from the provided passages. Without the requisite cultural knowledge, identifying the appropriate passage becomes challenging, even when given access to the entire set of passages.

Q (EN) : What is the name of this traditional Korean house, and during which dynasty did this architectural style become prominent?
A (EN) : This is a Hanok, and this architectural style became prominent during the Joseon Dynasty.
Keywords (EN) : Hanok, Joseon Dynasty

Q (ZH) : 这种传统的韩国房屋叫什么名字，这种建筑风格在什么朝代变得突出？
A (ZH) : 这是韩屋，这种建筑风格在朝鲜王朝时期变得突出。
Keywords (ZH) : 韩屋, 朝鲜王朝

Q (KO) : 이 전통 한국 가옥의 이름은 무엇이며, 어떤 왕조 때 이 건축 양식이 두드러지게 되었나요?
A (KO) : 이 가옥은 한옥이며, 이 건축 양식은 조선 왕조 때 두드러지게 되었습니다.
Keywords (KO) : 한옥, 조선 왕조

Passages

1) Hanok houses are known for their use of natural materials such as wood, clay, and paper, which help regulate temperature and humidity.
2) The Joseon Dynasty, lasting from 1392 to 1897, is known for its significant contributions to Korean culture, including advancements in architecture, art, and literature. The Hanok style flourished during this period.
3) The Great Wall of China, one of the most famous landmarks in the world, was built to protect Chinese states and empires against various nomadic groups from the north.
4) Traditional Korean houses often include features such as ondol (underfloor heating) and maru (wooden verandas), which are designed to suit the Korean climate.
5) Hanok is a traditional Korean house that is characterized by its unique architectural style, including tiled roofs and wooden beams. This style of housing became prominent during the Joseon Dynasty.

1) 传统的韩国房屋通常包括地暖（地板采暖）和木制走廊（木制阳台）等特点，这些设计是为了适应韩国的气候。
2) 韩屋是传统的韩国房屋，其特点是独特的建筑风格，包括瓦屋顶和木梁。这种住房风格在朝鲜王朝时期变得突出。
3) 中国的长城是世界上最著名的地标之一，建造它是为了保护中国的各个国家和帝国免受来自北方各种游牧部落的侵袭。
4) 韩屋以其使用天然材料如木材、黏土和纸张而闻名，这有助于调节室内温度和湿度。
5) 朝鲜王朝从1392年到1897年，因其对韩国文化的重要贡献而闻名，包括在建筑、艺术和文学方面的进步。韩屋风格在此期间繁荣发展。

1) 한옥은 기와 지붕과 나무 들보 등 독특한 건축 양식으로 특징지어지는 전통 한국 가옥이다. 이 주택 양식은 조선 시대에 두드러지게 나타났다.
2) 세계에서 가장 유명한 랜드마크 중 하나인 만리장성은 북쪽의 다양한 유목민 집단에 맞서 중국의 여러 국가와 제국을 보호하기 위해 건설되었다.
3) 1392년부터 1897년까지 이어진 조선 왕조는 건축, 예술, 문학 등 한국 문화에 중요한 기여를 한 것으로 알려져 있다. 한옥 양식은 이 시기에 번성했다.
4) 한옥은 실내 온도와 습도를 조절하는 데 도움이 되는 나무, 점토, 종이 등 천연 재료를 사용하는 것으로 유명하다.
5) 전통 한국 가옥에는 온돌(바닥 난방)과 마루(목조 베란다)와 같은 특징이 자주 포함되어 있으며, 이는 한국의 기후에 적합하게 설계되었다.

Figure 4: Examples of the created VLR-Bench data. (Korean culture)

Q (EN) : What is the name of the building in this image, and during which emperor's reign was it originally constructed?
A (EN) : The building is Hagia Sophia, and it was originally constructed during the reign of Emperor Justinian I.
Keywords (EN) : Hagia Sophia, Emperor Justinian I

Q (ZH) : 这张图片中的建筑叫什么名字，它最初是在哪个皇帝统治期间建造的？
A (ZH) : 这座建筑叫圣索菲亚大教堂，它最初是在查士丁尼一世皇帝统治期间建造的。
Keywords (ZH) : 圣索菲亚大教堂, 查士丁尼一世

Q (KO) : 이 이미지에 있는 건물의 이름은 무엇이며, 어느 황제의 치세 동안 처음 건설되었나요?
A (KO) : 이 건물은 아야 소피아이며, 유스티니아누스 1세 황제의 치세 동안 처음 건설되었습니다.
Keywords (KO) : 아야 소피아, 유스티니아누스 1세

Passages

1) Emperor Justinian I ruled the Byzantine Empire from 527 to 565 AD. He is known for his ambitious building projects, including the construction of the Hagia Sophia, which was intended to be the world's grandest church.
2) Hagia Sophia has served various roles throughout history, including as a mosque after the Ottoman conquest of Constantinople in 1453 and later as a museum in the 20th century.
3) The Grand Bazaar in Istanbul, one of the oldest and largest covered markets in the world, has been a major trading center since the 15th century and attracts millions of visitors annually.
4) Hagia Sophia, located in Istanbul, Turkey, was originally constructed as a Christian cathedral under the orders of Byzantine Emperor Justinian I. The construction started in 532 AD and was completed in 537 AD.
5) The architectural design of Hagia Sophia includes a massive dome, which was considered an engineering marvel of its time and influenced the development of architecture in both the Eastern Orthodox and Islamic worlds.

1) 圣索菲亚大教堂位于土耳其伊斯坦布尔，最初是在拜占庭皇帝查士丁尼一世的命令下建造的基督教大教堂。建筑始于公元532年，并于公元537年完工。
2) 伊斯坦布尔的大巴扎尔是世界上最古老和最大的室内市场之一，自15世纪以来一直是主要的贸易中心，每年吸引数百万游客。
3) 查士丁尼一世皇帝于公元527年至565年统治拜占庭帝国。他因其雄心勃勃的建筑项目而闻名，其中包括旨在成为世界上最宏伟教堂的圣索菲亚大教堂。
4) 圣索菲亚大教堂的建筑设计包括一个巨大的圆顶，这在当时被认为是工程奇迹，并影响了东正教和伊斯兰世界的建筑发展。
5) 圣索菲亚大教堂在历史上扮演了各种角色，包括在1453年奥斯曼帝国征服君士坦丁堡后作为清真寺使用，后来在20世纪作为博物馆。

1) 아야 소피아의 건축 설계에는 당시의 공학적 경이로 여겨졌던 거대한 돔이 포함되어 있으며, 동방 정교회와 이슬람 세계의 건축 발전에 영향을 미쳤습니다.
2) 터키 이스탄불에 위치한 아야 소피아는 비잔틴 황제 유스티니아누스 1세의 명령에 따라 원래 기독교 대성당으로 건설되었습니다. 건설은 서기 532년에 시작되어 537년에 완료되었습니다.
3) 유스티니아누스 1세 황제는 서기 527년부터 565년까지 비잔틴 제국을 통치했습니다. 그는 세계에서 가장 웅장한 교회를 건설하려는 의도로 아야 소피아를 포함한 야심 찬 건축 프로젝트로 잘 알려져 있습니다.
4) 아야 소피아는 역사적으로 다양한 역할을 수행해 왔으며, 1453년 오스만 제국이 콘스탄티노플을 정복한 후 모스크로 사용되었고, 20세기에는 박물관으로 사용되었습니다.
5) 이스탄불의 그랜드 바자르는 세계에서 가장 오래되고 큰 실내 시장 중 하나로, 15세기부터 주요 무역 중심지였으며 매년 수백만 명의 방문객을 끌어들입니다.

Figure 5: Examples of the created VLR-Bench data. (commonsense knowledge)

A.2 VLR-BENCH Construction Process

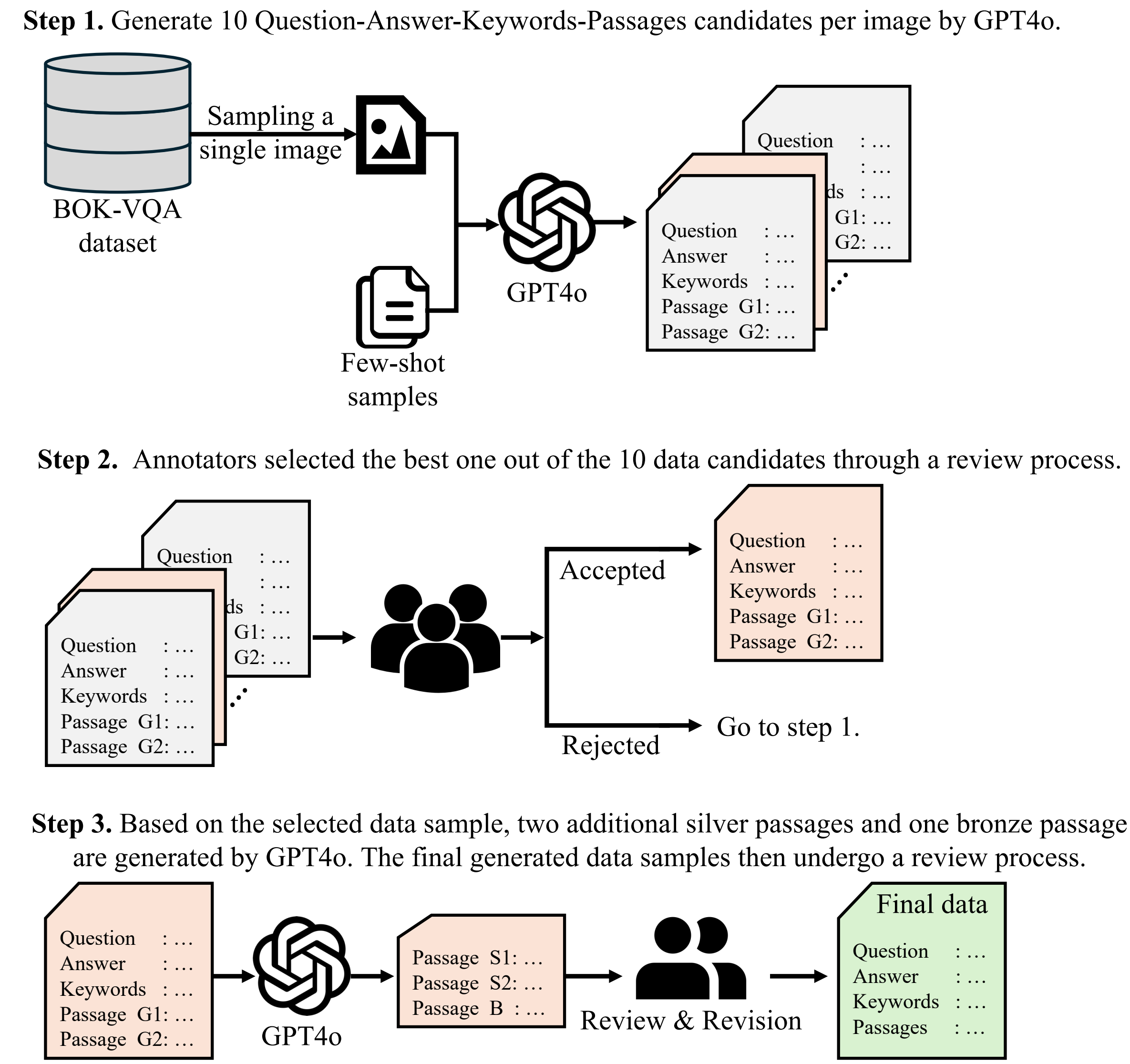


Figure 6: Overview of the VLR-BENCH dataset construction process.

Overview of Data Construction Procedure The image samples used in the dataset are sourced from the BOK-VQA (Kim et al., 2024) dataset, ensuring a wide range of visual content. The construction process involves few-shot learning and initial generation, annotator review and selection, passage expansion, and final review. GPT-4o generates candidate question-answer-passage sets based on few-shot examples, which are then reviewed and selected by human annotators. The selected sets are further expanded by GPT4o to create additional silver and bronze passages. The final dataset comprises a query, an answer, five passages (two gold, two silver, and one bronze), and two answer keywords for each image. Through a rigorous review process, the dataset maintains a high level of quality and relevance.

Annotation Guidelines To ensure the production and verification of high-quality data, we employed three computer science students. The annotators, aged 23, 23, and 27, included native speakers of Chinese and Korean, who were responsible for data in their respective languages. To generate data optimized for model training, we adhered to the guidelines for long-form sentences provided by BOK-VQA. However, when determining the Gold, Silver, and Bronze status of external knowledge, which is not covered by BOK-VQA, the annotators used their personal judgment. We proposed a maximum sentence length of 200 tokens for external knowledge. In cases where there were discrepancies in the corrections among annotators, discussions were held to revise in a more natural direction. Specifically, during the final data construction, there were many conflicts in selecting two keywords depending on the annotator's preferences. Therefore, a 27-year-old annotator proficient in both Chinese and Korean made the final selection by choosing two keywords from all the ones that had been selected at least once.

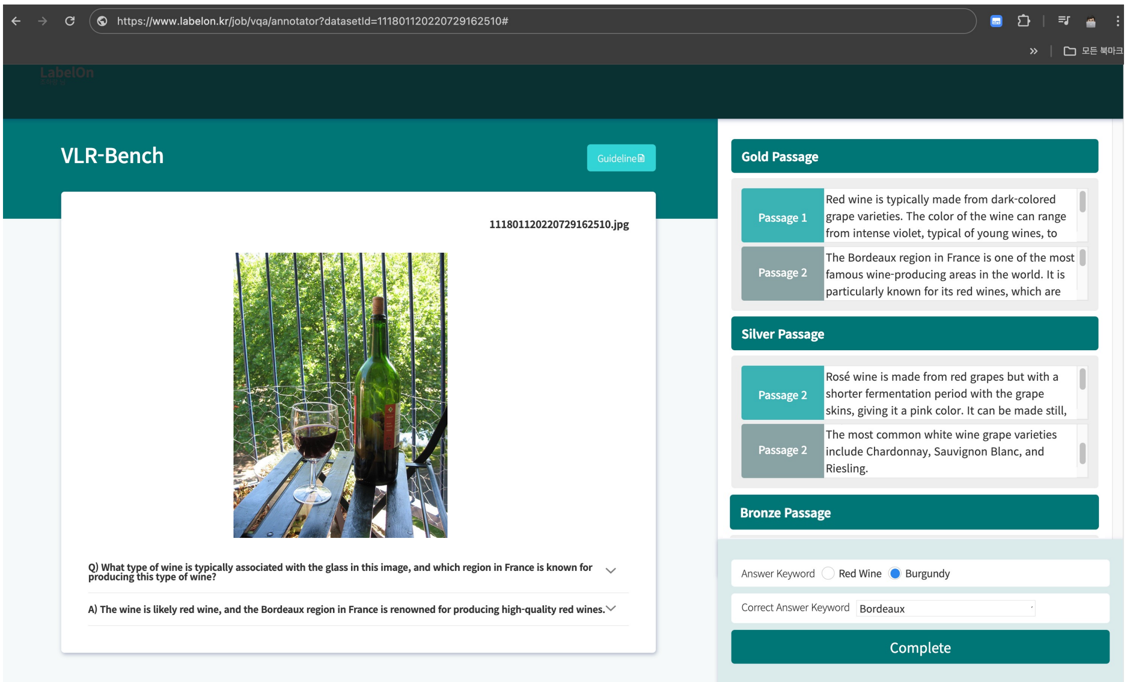


Figure 7: VLR-BENCH annotation tool.

A.3 VLR-Bench Few-shot Setup Examples

As mentioned in Subsection A.2, we generate 10 question-answer-keywords-passages candidates using few-shot samples. In this section, we demonstrate our few-shot examples.

Question(EN): Who is the architect that designed and directly oversaw the construction of this building, and in what architectural style was this cathedral designed?
Answer(EN): This building is the Sagrada Familia. The architect who designed and was responsible for the construction of the Sagrada Familia is Antoni Gaudí from Catalonia, Spain. He combined Gothic and Art Nouveau styles in his design.
Keywords(EN): Antoni Gaudí, Gothic and Art Nouveau styles

Question(ZH): 这座建筑物的设计师和负责建筑的建筑师是谁？这座大教堂是按照什么样式设计的？
Answer(ZH): 这座建筑是圣家堂。负责设计和建造圣家堂的建筑师是来自西班牙加泰罗尼亚的安东尼·高迪。他在设计中结合了哥特式和新艺术风格。
Keywords(ZH): 安东尼·高迪, 哥特式和新艺术风格

Question(KO): 이 건축물을 설계하고 직접 건축을 책임진 건축가는 누구이며, 이 성당은 어떤 양식으로 설계되었나요?
Answer(KO): 이 건축물은 사그라다 파밀리아 성당입니다. 사그라다 파밀리아 성당을 설계하고 건축을 책임진 건축가는 스페인 카탈루냐 출신의 안토니오 가우디입니다. 그는 고딕 건축 양식과 아르누보 양식을 결합하여 이 성당을 설계했습니다.
Keywords(KO): 안토니오 가우디, 고딕 건축 양식과 아르누보 양식

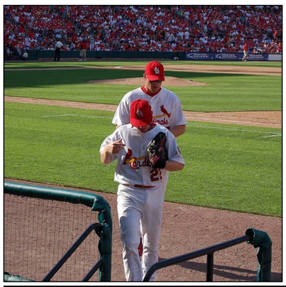
External-Knowledge:
1) he Basílica i Temple Expiatori de la Sagrada Família, otherwise known as Sagrada Família, is a church under construction in the Eixample district of Barcelona, Catalonia, Spain. It is the largest unfinished Catholic church in the world. Designed by Catalan architect Antoni Gaudí (1852–1926), in 2005 his work on Sagrada Família was added to an existing (1984) UNESCO World Heritage Site, "Works of Antoni Gaudí".On 7 November 2010, Pope Benedict XVI consecrated the church and proclaimed it a minor basilica.
2) On 19 March 1882, construction of Sagrada Família began under architect Francisco de Paula del Villar. In 1883, when Villar resigned, Gaudí took over as chief architect, transforming the project with his architectural and engineering style, combining Gothic and curvilinear Art Nouveau forms. Gaudí devoted the remainder of his life to the project, and he is buried in the church's crypt. At the time of his death in 1926, less than a quarter of the project was complete.

Figure 8: Examples of the few-shot sample.

B VLR-IF

B.1 VLR-IF Example

The following figures are examples from VLR-IF. Each example consists of a question, an answer, and a passage.



Q (EN) : What is the name of the baseball team represented by the players in the photo?
A (EN) : The baseball team represented by the players in the photo is the St. Louis Cardinals. The St. Louis Cardinals are a professional baseball team based in St. Louis, Missouri. They are a member of the National League Central Division in Major League Baseball (MLB).

Q (ZH) : 照片中的球员代表的棒球队名字是什么？
A (ZH) : 照片中的球员代表的棒球队是圣路易斯红雀队。圣路易斯红雀队是一支位于密苏里州圣路易斯的职业棒球队，属于美国职棒大联盟 国家联盟中区。

Q (KO) : 사진 속 선수들이 소속된 야구팀의 이름은 무엇인가요?
A (KO) : 사진 속 선수들이 소속된 야구팀은 세인트루이스 카디널스입니다. 세인트루이스 카디널스는 미주리주 세인트루이스를 연고로 하는 프로 야구팀입니다. 이 팀은 메이저 리그 베이스볼(MLB) 내셔널 리그 센트럴 디비전에 속해 있습니다

Passage

(EN) The St. Louis Cardinals are a professional baseball team based in St. Louis, Missouri. They compete in Major League Baseball (MLB) as a member club of the National League (NL) Central division. The team was established in 1882 and has won numerous World Series titles and National League pennants, making them one of the most successful and storied franchises in MLB history. The Cardinals' home games are played at Busch Stadium, which is located in downtown St. Louis. The team is known for its passionate fan base and has a long-standing rivalry with the Chicago Cubs.

(ZH) 圣路易斯红雀队是位于密苏里州圣路易斯的一支职业棒球队。他们作为国家联盟（NL）中区的成员俱乐部参加美国职业棒球大联盟（MLB）的比赛。该队成立于1882年，赢得了无数的世界大赛冠军和国家联盟锦旗，使他们成为MLB历史上最成功和最有故事的球队之一。 红雀队的主场比赛在位于圣路易斯市中心的布什体育场进行。该队以其热情的球迷群体而闻名，并与芝加哥小熊队有着长期的竞争关系。

(KO) 세인트루이스 카디널스는 미주리주 세인트루이스를 연고로 하는 프로 야구 팀입니다. 이 팀은 메이저 리그 베이스볼(MLB) 내셔널 리그(NL) 센트럴 디비전의 일원으로 경쟁합니다. 1882년에 설립된 이 팀은 수많은 월드 시리즈 타이틀과 내셔널 리그 페넌트를 획득하여 MLB 역사상 가장 성공적이고 전통 있는 프랜차이즈 중 하나로 자리매김했습니다. 카디널스의 홈 경기는 세인트루이스 다운타운에 위치한 부시 스타디움에서 열립니다. 이 팀은 열정적인 팬층으로 유명하며 시카고 컵스와 오랜 라이벌 관계를 유지하고 있습니다.

Figure 9: First example of the created VLR-IF data.

Q (EN) : What is the historical significance of the building shown in the photo?
A (EN) : The building in the photo is the Sándor Palace (Sándor-palota) located in Budapest, Hungary. It serves as the official residence of the President of Hungary. Originally built in 1806, it has historical significance as a symbol of Hungarian statehood and has housed various political leaders and dignitaries over the years.

Q (ZH) : 照片中建筑物的历史意义是什么？
A (ZH) : 照片中的建筑是位于匈牙利布达佩斯的山多尔宫（Sándor-palota）。它是匈牙利总统的官邸。该建筑始建于1806年，作为匈牙利国家象征具有历史意义，历年来接待了多位政治领导人和贵宾。

Q (KO) : 사진에 보이는 건물의 역사적 의미는 무엇인가요?
A (KO) : 사진에 보이는 건물은 헝가리 부다페스트에 위치한 산도르 궁전(Sándor-palota)입니다. 이 건물은 헝가리 대통령의 공식 거처로 사용됩니다. 1806년에 지어진 이 건물은 헝가리 국가의 상징으로서 역사적 의미를 지니며, 여러 정치 지도자와 귀빈들이 거쳐간 장소입니다.

Passage

(EN) The Sándor Palace (Sándor-palota) is located in Budapest, Hungary, near the Buda Castle. It was originally built in 1806 and has served various roles throughout its history. Since 2003, it has been the official residence of the President of Hungary. The palace is named after Count Vincent Sándor, who commissioned its construction. Over the years, it has been used for various governmental functions and has hosted numerous dignitaries. The building is an important symbol of Hungarian statehood and is located near other significant historical sites, such as the Buda Castle and the Hungarian Parliament Building.

(ZH) 桑多尔宫（Sándor-palota）位于匈牙利布达佩斯，靠近布达城堡。它最初建于1806年，在其历史上曾担任过各种角色。自2003年以来，它一直是匈牙利总统的官邸。 这座宫殿以委托建造它的文森特·桑多尔伯爵命名。多年来，它一直用于各种政府职能，并接待了众多贵宾。 这座建筑是匈牙利国家的重要象征，位于布达城堡和匈牙利国会大厦等其他重要历史遗址附近。

(KO) 샨도르 궁전(샨도르 팔로타)은 헝가리 부다페스트의 부다 성 근처에 위치해 있습니다. 이 궁전은 원래 1806년에 지어졌으며 역사적으로 다양한 역할을 해왔습니다. 2003년 이후로는 헝가리 대통령의 공식 거주지로 사용되고 있습니다. 궁전의 이름은 건축을 의뢰한 빈센트 샨도르 백작의 이름을 따서 지어졌습니다. 수년간 이곳은 다양한 정부 기능에 사용되었고 수많은 귀빈들을 맞이했습니다. 이 건물은 헝가리 국가의 중요한 상징이며, 부다 성과 헝가리 국회의사당과 같은 다른 중요한 역사적 장소 근처에 위치해 있습니다.

Figure 10: Second example of the created VLR-IF data.

B.2 VLR-IF Construction Process

Figure 11 illustrates the construction process of the VLR-IF dataset. The dataset consists of 9K images, with each image corresponding to a single query, answer, and passage. Following the approach used in building VLR-Bench, we provided GPT-4o with few-shot samples (including image, query, answer, and passage) along with the image example we wanted to generate. Then, we generated the query, answer, and passage for the image example. To enhance the model's ability to select valid passages, we determined that it would be desirable to include diverse passages for each image. Accordingly, we assumed the

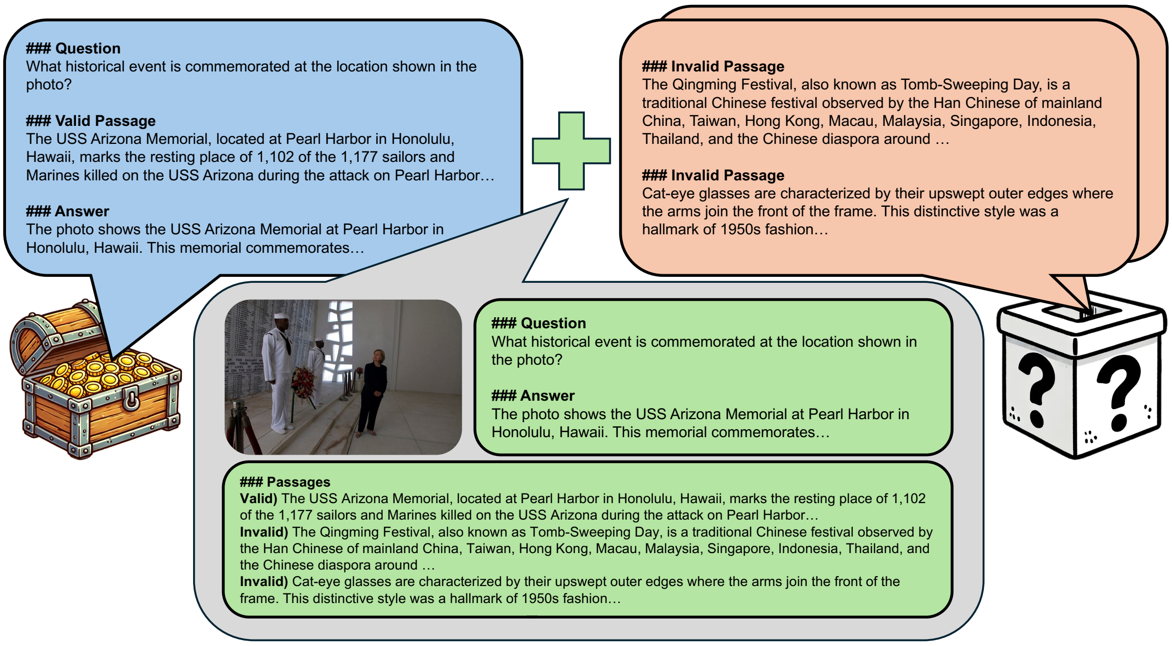


Figure 11: The process of constructing the VLR-IF dataset.

original passage of each image to be a valid passage and randomly extracted passages from other images to set them as invalid passages. When only invalid passages are used, the model is designed to generate the following response: “The provided knowledge does not pertain to the image, so I can’t answer the question.” Ultimately, we constructed a total of 32,000 datasets by combining valid and invalid passages in the following manner: $\{V\}$, $\{I\}$, $\{V, I\}$, $\{V, I, I\}$.

- $\{V\}$: Only the valid passage, 9,000 datasets.
- $\{I\}$: Only one invalid passage, 5,000 datasets. In this case, the training data was constructed to output "Insufficient search results found, making inference impossible" when encountering such instances.
- $\{V, I\}$: One valid and one invalid passage, 9,000 datasets.
- $\{V, I, I\}$: One valid and two invalid passages, 9,000 datasets.

C Details of Experimental Environments

C.1 Baseline Models

LMM	LLM	#VIT	source	latest update(dd.mm.yyyy)
LLAVA1.5	Llama2-13 B	665 K	liuhaotian/llava-v1.5-13b	10.05.2024
LLAVA-LLAMA-3	Llama3-8 B	1.2 M	xtuner/llava-llama-3-8b-v1_1-transformers	28.04.2024
X-LLAVA	Llama2-13 B	407 K	MLP-KTLim/X-LLaVA	02.01.2024
QWEN-VL	Qwen 7 B	350 K	Qwen/Qwen-VL-Chat	26.01.2024

Table 4: The Vision-Language Models (VLMs) used for evaluation on VLR-BENCH were accessed through the Hugging Face Transformers library version 4.32.0 (Wolf et al., 2020)

LLAVA-LLAMA-3. The LLaVA-based model fine-tuned from meta-llama/Meta-Llama-3-8B-Instruct[3] and CLIP-ViT-Large-patch14-336[4] with ShareGPT4V-PT and InternVL-SFT by XTuner.

X-LLAVA. We selected X-LLaVA, a Korean and English Multimodal LLM, as the base model for the Korean and English benchmarks. X-LLaVA was trained on a dataset of 91K English-Korean-Chinese

[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[4]https://huggingface.co/openai/clip-vit-large-patch14-336

multilingual and multimodal learning data.

QWEN-VL-CHAT. We employed Qwen-7B as the LLM and Openclip ViT-bigG as the Visual Encoder. The Qwen-VL model is constructed by connecting the LLM and Visual Encoder to a randomly initialized cross-attention layer. Finally, Qwen-VL-Chat is a model obtained by fine-tuning Qwen-VL using an instruction-following dataset.

C.2 Hyperparameter Settings

	value
Optimizer	AdamW
learning_rate	5.0e-5
Dropout	0.05
lr_scheduler	cosine
Epoch for IT	1
Epoch for PT	1
sequence_len	4096
Batch size	1
Random Seed	1004
llm	lora
Low-rank size	64
lora_alpha	128
lora_dropout	0.05
lora_trainable	q, v, k, o, gate, down, up_proj
LoRA layer	q, k, v

Table 5: Applied hyperparameter settings.

The hyperparameter settings used in this study can be found in Table 5. Models utilizing LoRA were trained using only a portion of the attention layers indicated in the table, as well as θ^e and θ^h, and the size of the low-rank matrices was set to 64. All models were trained for 1 epoch.

Experiment Reproduction. We are making the training code, trained models, and data used for testing available to allow for exact reproduction of the experiments conducted in this study. The qualitative responses generated by the models during the experiments can be downloaded from the following site, with files named after the models corresponding to the experimental results of those models.

C.3 Performance Comparison on Various Passage Types.

Model	English					
	PSG	MS	R-2	R-L	BLEU	BERT-Score
LLAVA-LLAMA-3	GG	**84.8**	**39.9**	**50.2**	**20.6**	**84.1**
	G	58.4	30.9	41.4	14.8	81.2
	GS	68.0	33.5	44.3	16.2	82.0
	GB	59.6	30.4	40.9	14.6	81.0
	SS	41.2	23.3	33.7	10.6	78.3
	SB	38.4	23.4	33.5	10.4	78.3
	B	22.4	19.5	29.6	9.2	76.4
LLAVA-LLAMA-3+VLR-IF	GG	**86.4**	**49.5**	**62.4**	**34.0**	**87.3**
	G	68.0	42.6	56.1	26.1	85.1
	GS	73.6	42.1	55.9	26.2	85.0
	GB	69.2	41.6	55.3	26.9	84.8
	SS	48.0	33.3	47.1	18.5	81.9
	SB	46.0	33.9	47.9	19.6	82.1
	B	22.4	22.0	35.0	15.0	76.7

Table 6: Performance Comparison of LLaVA-LLaMA-3 with and without VLR-IF Training on Various Passage Types. The results demonstrate that, regardless of the passage type, the model trained on VLR-IF consistently outperforms its counterpart without VLR-IF training across all evaluation metrics. This finding supports the hypothesis that the VLR-IF dataset effectively enhances the model's ability to select crucial information from passages, enabling it to better follow user instructions based on the given image.

Model	InfoSeek
Llava-Llama-3	42.9
Llava-Llama-3+VLR-IF(EN)	44.5

Table 7: Performance difference in InfoSeek depending on VLR-IF training when using a search engine as a passage retriever.

Table 7 presents the results of evaluating the InfoSeek benchmark performance with and without VLR-IF training using the LLAVA-LLAMA-3 model. VLR-IF was trained solely on the English dataset, and the Oracle was used as the Retriever model. The evaluation showed that the model trained with VLR-IF achieved a 2.6 points improvement in performance even on the external benchmark dataset, InfoSeek.

D Comprehensive Analysis of Datasets

D.1 VLR-Bench Validity Analysis

Lang.	Model	VLR-IF			Quantitative Avg.	GPT4o Score
		EN	ZH	KO		
EN	Llava1.5 (Liu et al., 2024)	✗	✗	✗	48.5	9.11
	Llava-Llama-3	✗	✗	✗	48.2	9.03
	Llava-Llama-3+VLR-IF(En)	✓	✗	✗	52.9	**9.40**
	X-Llava (Shin et al., 2024)	✗	✗	✗	49.5	9.10
	X-Llava+VLR-IF(En)	✓	✗	✗	51.4	9.27
	X-Llava+VLR-IF(En+Ko)	✓	✗	✓	52.1	9.28
	Qwen-VL-Chat	✗	✗	✗	**53.5**	9.30
ZH	QWEN-VL-CHAT (Bai et al., 2023)	✗	✗	✗	60.3	8.33
	QWEN-VL-CHAT+VLR-IF(Zh)	✗	✓	✗	**64.8**	**9.26**
KO	X-Llava	✗	✗	✗	43.1	7.62
	X-Llava+VLR-IF(Ko)	✗	✗	✓	**50.0**	8.35
	X-Llava+VLR-IF(En+Ko)	✓	✗	✓	49.8	**8.59**

Table 8: A table illustrates the results of the qualitative assessment using GPT-4o. Quantitative Avg. is the average result of the quantitative evaluation conducted Table 1.

To validate the quantitative evaluation results of VLR-Bench, we conducted a qualitative assessment using GPT-4o. GPT-4o was provided with images from the VLR-Bench dataset, queries, external knowledge required for answering, and the model's responses. Based on this information, the model's responses were evaluated on the following four aspects: (1) Assessment of the model's selection of Gold Passages and the use of Silver Passages. (2) Evaluation of the accuracy, completeness, and readability of the model's responses. (3) Verification of the model's fulfillment of the query requirements. (4) Examination of whether additional content from Silver Passages or Bronze Passages was included based on the length of the responses. This rigorous evaluation ensures the reliability and validity of the quantitative results obtained from VLR-Bench.

Additionally, GPT-4o outputs the evaluation scores along with the reasoning behind the evaluations. Through this process, we conducted a reliable qualitative assessment and confirmed that the results exhibited a distribution similar to the quantitative evaluation results of VLR-Bench, as shown in Table 8. The Figure 12 illustrates the prompt and responses provided to GPT-4o for the qualitative assessment.

D.2 VLR-IF Validity Analysis

To investigate the impact of the VLR-IF dataset, we evaluated its effect on passage selection in our experiments. We chose English as the target language for the experiments and used the Llava-Llama-3 model, which received the highest evaluation in this language. The experiment proceeds as follows: first, we provide the model with passages, an instruction, and an image. The model then selects two passages necessary to follow the instruction related to the image. As shown in the Table 9, the model fine-tuned on VLR-IF demonstrates a substantial improvement of 26.0 and 25.1 points in EM and F1 scores, respectively, compared to the model without VLR-IF training. The results suggest that the VLR-IF dataset can enhance the ability to select the necessary passages based on images and queries.

Model	EM	F1
LLAVA-LLAMA-3	2.0	15.9
LLAVA-LLAMA-3+VLR-IF(EN)	28.0	41.0

Table 9: Passage Selection Performance with and without VLR-IF Training. EM is the exact matching score, while F1 is the harmonic mean of precision and recall.

System Prompt
You need help with the following question involving an image.

The model will analyze the image and instructions, referring to five passages to provide an answer. Two of these passages contain essential information directly related to the image and the question, which we call the "Gold Passage." Two of the passages contain information related to the topic but are not central to answering the question; we call these "Silver Passages." The remaining passage is unrelated to the image and the question, which we call the "Bronze Passage." We give you the five passages.

You will meticulously evaluate the answers provided by the language model to the questions. To ensure the fairest evaluation, you must adhere to the following rules:

Basic Rules
1. Focus on how well the model references the Gold Passages to answer the question and how effectively it filters out the unnecessary Passages.
2. Focus on the accuracy, completeness, and readability of the answers.
3. Analyze in detail whether anything was missed from the question's requirements.
4. Do not let the length of the answer influence the evaluation.

If the answer violates these rules, apply a significant penalty to the score. Evaluation Output Format Provide a brief evaluation of the answer, and if it does not receive a 10.0, explain in 1-2 sentences why it did not. Then, score the answer on a scale of 1.0 to 10.0. Do not use symbols when expressing scores; for example, use "score: 4.0" instead of "4.0/10.0" or "5.0" Finally, follow the format below to generate your response.

Output Format
Evaluation: [Evaluation content]
Score: [Number]

Instruction
Question: What is the name of the pagoda in this image, and during which dynasty was it constructed?

Gold Passage1: Dabotap Pagoda is one of the two famous pagodas located at Bulguksa Temple in Gyeongju, …

Gold Passage2: The Unified Silla Dynasty, which lasted from 668 to 935 AD, is known for its significant …

Silver Passage1: Gyeongju, the capital of the Unified Silla Dynasty, is often referred to as …

Silver Passage2: The Dabotap Pagoda stands at 10.4 meters tall and is known for its intricate …

Bronze Passage: Bulguksa Temple, where the Dabotap Pagoda is located, is a UNESCO …

Model Answer: The pagoda is the Dabotap Pagoda, and it was constructed during the Unified Silla Dynasty in the 8th century.

Evaluation: The model correctly identifies the glass in the image as a champagne coupe and accurately describes its traditional use for serving champagne and sparkling wines. It also mentions the historical design inspiration and the shift in popularity to the champagne flute, which is relevant and accurate. The model effectively uses information from both Gold Passages and appropriately filters out the Silver and Bronze Passages. The answer is complete, accurate, and readable.

Score: 10.0

Figure 12: Examples of prompts used with GPT models in qualitative evaluations.

D.3 Related Datasets

Dataset	Image Source	# of Instances	Multilingual	Parallel	Open	Qualitative	Quantitative	With Passages
K-VQA	Wikipedia	183 K	✗	✗	✓	✗	✓	✗
S3VQA	Open Images Dataset	6 K	✗	✗	✓	✗	✓	✗
OK-VQA	COCO	14 K	✗	✗	✓	✗	✓	✗
A-OKVQA	COCO	24 K	✗	✗	✓	✗	✓	✗
ViQuAE	Wikipedia	3 K	✗	✗	✓	✗	✓	✗
OVEN-Wiki	Wikipedia	5,8 M	✗	✗	✓	✗	✓	✗
InfoSeek	Wikipedia	1.36 M	✗	✗	✓	✗	✓	✗
Encyclopedic VQA	iNaturalist, Google Landmarks Dataset	1.036 M	✗	✗	✓	✗	✓	✗
Ours	BOK-VQA	32.3 K	✓	✓	✓	✓	✓	✓

Table 10: Summary of the multimodal VQA (Visual Question Answering) benchmark dataset. 'Parallel' indicates that the dataset can be used for translation tasks. 'Qualitative' refers to the availability for quantitative evaluation, while 'Quantitative' refers to the availability for qualitative evaluation. 'With Passages' denotes whether passages are provided in the benchmark dataset.

Table 10 provides information on the size and domains of major VLM evaluation datasets that utilize external knowledge. The VLR-BENCH dataset proposed in this study is structurally similar to the Encyclopedic VQA dataset, which includes test data containing 1,000 gold passages. However, VLR-BENCH differs in two key ways: (1) instead of a single gold passage, each query is paired with five passages—two `Gold Passages`, two `Silver Passages`, and one `Bronze Passage`, and (2) it consists of parallel corpora in English, Chinese, and Korean, making the test data more than four times larger, even though the total number of samples is smaller. Moreover, unlike the automatically generated Encyclopedic VQA, all passages in VLR-BENCH have been manually reviewed, with a strong emphasis on quality control. By categorizing passages into gold, silver, and bronze, models must distinguish between useful and less relevant information to generate accurate answers. This design allows for a more nuanced evaluation of how well a VLM can utilize gold passages while avoiding the silver and bronze ones from the top-k retrieved results, setting VLR-BENCH apart from existing datasets.

D.4 Correlation analysis between Passages, Ground Truth, and Questions.

Lang.	Passage-type	Bert Score f1	Rouge-1	Rouge-2	Rouge-L
Passages & Ground-truth output					
EN	Gold	**78.04**	**44.96**	**20.98**	**31.92**
	Silver	72.11	25.59	5.978	18.92
	Bronze	67.81	22.10	2.299	17.04
ZH	Gold	**77.08**	**20.78**	**7.16**	**20.50**
	Silver	70.49	4.15	1.08	4.15
	Bronze	66.28	0.38	0.0	0.38
KO	Gold	**77.20**	**19.05**	**6.90**	**18.78**
	Silver	71.98	3.47	0.77	3.47
	Bronze	66.48	0.38	0.0	0.38
Passages & Questions					
EN	Gold	**66.55**	**28.42**	**4.67**	**18.93**
	Silver	65.70	21.48	1.97	15.79
	Bronze	64.48	24.11	2.74	17.26
ZH	Gold	**67.13**	**1.2**	0.0	**1.2**
	Silver	65.85	0.2	0.0	0.2
	Bronze	63.70	0.0	0.0	0.0
KO	Gold	**68.15**	**1.2**	0.0	**1.2**
	Silver	67.26	0.2	0.0	0.2
	Bronze	66.28	0.380	0.0	0.380

Table 11: Examining the correlation between Passages and GT reveals that, irrespective of the language used, the correlations are ordered in the sequence of Gold, Silver, and Bronze. This suggests that to successfully perform VLR-BENCH, it is necessary to appropriately utilize the Gold Passage. Meanwhile, investigating the correlation between Passages and Questions indicates that the level of correlation remains consistent across various types of Passages. These results demonstrate that Questions alone are insufficient for successfully completing VLR-BENCH, and that both images and Passages must be utilized together.