A Simple Data Augmentation Strategy for Text-in-Image Scientific VQA

Belal Shoer, Yova Kementchedjhieva
MBZUAI
{belal.shoer,yova.kementchedjhieva}@mbzuai.ac.ae

Abstract

Scientific visual question answering poses significant challenges for vision-language models due to the complexity of scientific figures and their multimodal context. Traditional approaches treat the figure and accompanying text (e.g., questions and answer options) as separate inputs. EXAMS-V introduced a new paradigm by embedding both visual and textual content into a single image. However, even state-of-the-art proprietary models perform poorly on this setup in zero-shot settings, underscoring the need for task-specific fine-tuning. To address the scarcity of training data in this "text-in-image" format, we synthesize a new dataset by converting existing separate image-text pairs into unified images. Fine-tuning a small multilingual multimodal model on a mix of our synthetic data and EXAMS-V yields notable gains across 13 languages, demonstrating strong average improvements and cross-lingual transfer.[1]

1 Introduction

Vision-language models (VLMs) have advanced AI by enabling multimodal reasoning, facilitating more natural user interaction in tasks such as Visual Question Answering (VQA) and captioning. Antol et al. (2015) proposed VQA as a task that spans language and image to generate an accurate response. The VQA task has evolved rapidly with applications and benchmarks in domains such as science (Lu et al., 2022), chart understanding (Masry et al., 2022), document analysis (Mathew et al., 2020), medical imaging (Hasan et al., 2018), and other real-world applications. VQA tasks typically follow either a multiple-choice or open-ended format.

In multiple-choice scientific VQA, the input typically consists of an image (figure, table, chart) accompanied by a question and answer choices in text form. The task requires reasoning over both image and text components to select the correct answer. These modalities are often processed separately in multimodal models. However, in practice, the text is often embedded within the visual modality, for example, screenshots of digital exams or textbook photos. To address this, Das et al. (2024) introduced a new scientific VQA benchmark that consists of images with embedded questions, providing a robust benchmark for evaluating model performance under realistic conditions. The EXAMS-V dataset includes two splits: train (16.5K instances) and test (4.8K instances), spanning 15 languages.

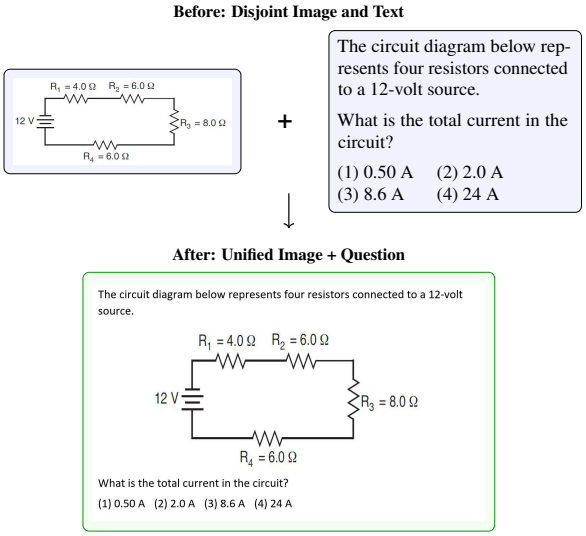

Figure 1: Synthetic data generation via mapping of a disjoint figure and text into a unified image.

This text-in-image formulation of scientific VQA either requires a separate Optical Character Recognition (OCR) step, which may introduce noise, or, preferably, VLMs with strong inherent OCR capabilities that can jointly reason over visual content and embedded text. Yet, current VLMs typically benefit from text as opposed to text-in-image format, even if the text is just describing the contents of the image itself (Vineet et al., 2024). EXAMS-V approached the task in a zero-shot man-

[1]Dataset: https://huggingface.co/datasets/Shoir/Scientific_VQA

Proceedings of the 9th Widening NLP Workshop, pages 100–105

ner, without leveraging the training data. Fine-tuning for reasoning over visually-embedded text is likely to improve performance.

While we already have the EXAMS-V training split in the text-in-image format, we find that it provides limited coverage, with an average of 1,415 data points per language. To address this, we augment the training set by synthesizing text-in-image data points derived from disjoint scientific VQA datasets, in four languages: Chinese, English, Italian, and German. This results in approximately 1,742 additional examples on average per language.

2 Background

In this section, we provide the necessary background for our work. We begin by reviewing traditional datasets and benchmarks commonly used in scientific VQA, highlighting their strengths and limitations. We then review EXAMS-V, the primary benchmark our work builds upon, and describe our chosen VLM, PaLIGemma (Steiner et al., 2024), explaining the rationale behind its selection.

2.1 Traditional Datasets

There are a number of multi-modal scientific VQA datasets that span multiple scientific fields, such as physics, chemistry, biology, mathematics, and geology. ScienceQA (Lu et al., 2022), was introduced as an English monolingual scientific multimodal dataset that has been collected from elementary and high school curricula. MMMU (Yue et al., 2024a) is another English-language scientific benchmark compiled from college exams and textbooks to challenge the VLMs' abilities on multi-modal multi-discipline subjects. These datasets treat vision and language as separate inputs, whereas EXAMS-V presents a novel approach by combining both modalities in a single image.

We harvest our data from 5 different datasets namely, M3EXAM (Zhang et al., 2023), CMMU (Zheqi He and Huang, 2024), M4U (Wang et al., 2024a), MMMU-PRO (Yue et al., 2024b), and Pinocchio (Federici, 2024). Since these datasets separate language and vision components, we synthetically combine the question and answer text with the corresponding figures to create text-in-image training examples.

Synthetic data generation has been shown to improve VLMs' performance. Chen et al. (2024a) generated a dataset of 1.3M examples and showed that small models can match or even outperform larger ones when trained on synthetic data. Moreover, Liu et al. (2024b) reported improved performance using their synthetic dataset.

2.2 Text-in-Image Datasets: EXAMS-V

The composition of language and vision poses a significant challenge to VLMs. Wang et al. (2024b) found that in spatial reasoning tasks, VLMs rarely outperform their traditional LLM counterparts and when provided with both image and text, they rely less on the visual modality.

EXAMS-V is a multilingual multimodal benchmark that consists of 20,932 multiple-choice questions curated from national exams from multiple nations. It contains two data formats: 15,846 text-only and 5,086 text-and-visual images.

2.3 PaliGemma

VLMs are widely adopted for their strong generalization across tasks such as image captioning, VQA, and visual grounding. Google recently introduced PaLIGemma 2, an enhanced version of PaLIGemma that integrates the more powerful Gemma 2 language model together with the SigLIP vision encoder. It supports three image resolutions: 224^2, 448^2, and 896^2 pixels.

PaLIGemma 2 was trained in stages: initially on 1 billion image-text pairs at 224^2 using the combined SigLIP So400m and Gemma 2 checkpoints, followed by 50 million examples at 448^2 and 10 million at 896^2, and finally on a mix of academic tasks including VQA, captioning, and detection.

We chose to fine-tune PaLIGemma 2 for three main reasons. First, it is lightweight, making it a practical alternative to large proprietary models. Second, it supports 34 languages, aligning well with our multilingual goals. Third, it offers flexibility in size and resolution and has been pre-trained on tasks relevant to our setting.

3 Data Augmentation

This section outlines our method for generating synthetic text-in-image instances for VQA and introduces the pre-trained VLM used in our experiments. We focus on the text-with-visual format rather than the text-only due to its limited presence in EXAMS-V, with only 5,162 such images.

We use data from five datasets that provide separate text and image pairs: M3EXAM, CMMU, M4U, MMMU-PRO, and Pinocchio. These datasets span multiple languages and subjects. We

Dataset	Languages	Used	Total
M3Exam	en (610), it (228), zh (351)	1,189	12,317
CMMU	zh	1,095	3,603
M4U	de	2,183	8,931
Pinocchio	it	1,392	136,849
MMMU-Pro	en	1,109	5,190
Total	**4**	**6,968**	**166,890**

Table 1: Number of questions used from each dataset compared to the total available questions.

focus on Chinese, English, Italian, and German. We filter the data to retain only science-related instances, primarily from Chemistry, Physics, Biology, Biochemistry, and Engineering. Each instance is formatted consistently, with the question at the top, followed by the figure and answer options. For an example of our method, refer to Figure 1.

To simulate realistic exam formats, Hanzi and Latin scripts are rendered using randomly selected fonts and dark text colors. We use common fonts such as SimSun and SimHei for Hanzi, and Arial and Times New Roman for Latin script. To reflect typical document formatting, text colors are sampled from a set of dark grayscale tones, with a strong bias toward black as detailed in Appendix B. We fix the random seed to 42 for reproducibility. To encourage generalization during fine-tuning, the option format (letters or numbers) is chosen uniformly at random for each synthetic instance.

4 Experiments

4.1 Experimental Setting

We fine-tune the PaliGemma 2-mix variant with 448^2 pixel input resolution. We freeze the vision encoder and the projection layer, training the language decoder for 5 epochs using AdamW with a learning rate of 2×10^{-5}, weight decay of 1×10^{-6}, batch size of 64, Eager attention, and a learning rate schedule with linear warm-up over the first 0.05% of the training steps followed by cosine decay.

To assess the utility of our synthetic text-in-image dataset, we fine-tuned two variants of the model under comparable training settings. The first variant is trained on a combination of the EXAMS-V training split and our synthetic data (FT-EV+SYN), while the second variant was trained exclusively on the original EXAMS-V training split without any synthetic augmentation (FT-EV).

We report results on the EXAMS-V test set, in term of accuracy of the multiple-choice answer that the model generates. As additional strong baselines, we include InternVL3-2B (Chen et al., 2024b) and LlaVA-Next (Mistral-7B) (Liu et al., 2024a).

4.2 Results

Main Results The main results are reported in Table 2, along with the number of train and test data points available for each language in the base dataset, EXAMS-V, as these values become relevant to the discussion below.

Both of our fine-tuned models, FT-EV and FT-EV+SYN, outperform the off-the-shelf PaliGemma 2 model (Non-FT). FT-EV+SYN achieves the highest average accuracy across the four augmented languages (zh, en, it, de) at 33.3%, outperforming FT-EV (32.4%), InternVL3-2B (28.7%), and LLaVA-NeXT (20.3%) by 0.9, 4.6, and 13.0 percentage points, respectively. It surpasses FT-EV in 3 out of the 4 languages, with the largest gain in German (+3.9 points). The only exception is Chinese, where performance slightly declines by 1.7 percentage points, possibly because this language is already well-represented in EXAMS-V (with 3308 train data points), reducing the benefit of additional synthetic data.

On average across all 13 languages, our targeted data augmentation leads to a slight decrease of 0.5 percentage points, possibly due to representational bias toward the augmented subset. Interestingly, several non-augmented languages show improvements, suggesting that synthetic data can enhance cross-lingual generalization. For example, Arabic improves by 1.7 points and Hungarian by 0.7 points over FT-EV. Slovak shows an even larger improvement of 13.0 points, but this may be influenced by the small number of test instances in Slovak (only 46), which can increase variance in performance estimates. Other languages also have limited test coverage; for example, Spanish and Polish each have only 100 test instances, which may explain the notable performance drop observed for FT-EV+SYN compared to FT-EV (Spanish: 67.0 to 59.0; Polish: 30.0 to 22.0).

Separate Modality Analysis EXAMS-V includes two image formats: text-only images and images containing both text and visuals (e.g. figures). Here, we investigate how our data augmentation affects each subset. As seen if Figure 2, both fine-tuned models, FT-EV and FT-EV+SYN, outperform the non-fine-tuned baseline across both formats. FT-EV+SYN achieves the highest accu-

Model	zh	en	it	de	hr	hu	ar	fr	pl	es	bg	sr	sk	Rel. avg.	Avg.
Train split	3308	1992	2571	2573	3207	3122	293	199	2285	190	1648	887	–	–	–
Test split	600	347	562	279	585	535	517	224	100	100	400	502	46	–	–
Non-FT	24.8	21.3	23.1	29.0	25.3	27.1	23.2	34.8	22.0	31.0	30.2	22.5	17.4	24.6	25.5
FT-EV	**32.5**	22.5	32.4	42.3	**32.3**	30.3	25.0	**47.8**	**30.0**	**67.0**	**32.5**	**27.9**	37.0	32.4	**35.3**
FT-EV+SYN	30.8	**23.6**	**32.6**	**46.2**	31.8	**31.0**	**26.7**	44.2	22.0	59.0	28.2	25.7	**50.0**	**33.3**	34.8
InternVL3-2B	27.8	21.3	33.8	35.5	27.4	27.1	14.3	47.8	24.0	43.0	21.3	29.9	19.6	29.6	28.7
LLaVA-NeXT	14.2	18.7	25.6	24.7	5.8	18.9	3.3	23.2	23.0	29.0	15.8	26.3	17.4	20.3	17.5

Table 2: Performance comparison of the non-fine-tuned PaliGemma model (**Non-FT**), fine-tuning on EXAMS-V (**FT-EV**), and fine-tuning on EXAMS-V with synthetic data (**FT-EV+SYN**). Results are shown alongside strong vision-language baselines (InternVL3-2B and LLaVA-NeXT-Mistral-7B). **Rel. avg.** is the average over zh, en, it, and de; **Avg.** is the overall average across all languages. The number of training and test instances for FT-EV+SYN is shown in the top. Bolded values indicate the best scores within the main results in the top three rows.

Model	zh	en	it	de	Avg.
FT-zh	**32.3**	23.3	28.1	31.5	28.8
FT-en	29.3	**25.4**	25.3	30.8	27.7
FT-it	28.5	18.7	26.7	34.1	27.0
FT-de	26.5	21.6	28.8	34.1	27.8
FT-EV+SYN	30.8	23.6	**32.6**	**46.2**	**33.3**

Table 3: Accuracy of language-specific vs. mixed-language fine-tuned models, evaluated per language. All of these models are trained with augmented data.

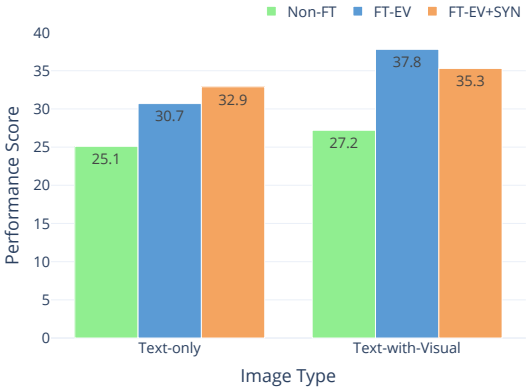

Figure 2: Performance comparison of fine-tuned and non-fine-tuned models across different image types, averaged over the four target languages.

racy on text-only images (32.9%), while FT-EV performs best on text-with-visual images (37.8%.) This is an unexpected finding, as it suggests that our synthetic data points, designed to contain both text and visuals, benefit text-only questions, but not questions containing both text and visuals.

Another interesting observation in Figure 2 is the relatively higher performance of all models on the text-with-visuals portion of the data, compared to text-only. Contrary to prior findings on the relative complexity of multimodal questions, here we see these questions emerging as easier for the models. It remains to be explored whether this is a property of the data or of the models.

Monolingual Training Having established the performance of our model under multilingual training, we now experiment with monolingual training to assess the extent of cross-lingual transfer or interference. As shown in Table 3, we fine-tune separate models on monolingual augmented subsets of the dataset and compare their accuracy to that of our multilingual model, FT-EV+SYN. The multilingual model demonstrates superior performance compared to its monolingual counterparts on average across the four languages, as well as in both German and Italian (by 12.1 and 5.9 percentage points, respectively). This considerable gap highlights the benefits of cross-lingual training.

We further observe an intriguing cross-lingual effect: the FT-zh and FT-it outperform the Italian monolingual model, FT-it, on the Italian test set, by a considerable margin of 1.5 to 2 points. This may be attributed to a distributional mismatch between the augmented Italian augmented data and the Italian instances in EXAMS-V.

5 Conclusion

In this study, we augment the EXAMS-V dataset with synthetic text-in-image instances for 4 languages. Our experiments demonstrate improved performance across the four languages on average, albeit on text-only questions and not on questions containing visuals. We find that multilingual fine-tuning outperforms monolingual fine-tuning on average, indicating a positive cross-lingual transfer.

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 (ICCV)*.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *Preprint*, arXiv:2402.11684.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*.

Edoardo Federici. 2024. Pinocchio: An italian, culture-aware, language understanding dataset.

Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, and Henning Müller. 2018. Overview of the ImageCLEF 2018 medical domain visual question answering task. In *CLEF2018 Working Notes*, CEUR Workshop Proceedings, Avignon, France. CEUR-WS.org <http://ceur-ws.org>.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. 2024b. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for VQA on document images. *CoRR*, abs/2007.00398.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, and 1 others. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Vibhav Vineet, Xin Wang, and Neel Joshi. 2024. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *2024 Neural Information Processing Systems*.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024a. M4u: Evaluating multilingual understanding and reasoning for large multimodal models.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024b. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models.

Pengfei Zhou Richeng Xuan Guang Liu Xi Yang Qiannan Zhu Zheqi He, Xinya Wu and Hua Huang. 2024. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning. *arXiv preprint arXiv:2401.14011*.

A Hyperparameters

Parameter	Value
Attention Implementation	Eager
Learning Rate	2e-5
Weight Decay	1e-6
Batch Size	64
Optimizer	adamw_torch
Scheduler	Warm-up + Cosine Decay
Epochs	5

Table 4: Fine-tuning hyperparameters.

B Data Specifications

Script	Fonts
Hanzi	Microsoft YaHei, SimSun, FangSong, SimHei, Alibaba PuHuiTi Regular
Latin	Arial, Times New Roman, Courier New, Verdana, Calibri

Table 5: Fonts used for generating Hanzi and Latin rendered instances.

RGB Value	Sampling Weight (%)
(0, 0, 0) - Black	90
(20, 20, 20)	2
(43, 43, 43)	2
(82, 82, 82)	2
(138, 138, 138)	2
(168, 168, 168)	2

Table 6: Grayscale RGB values used for text rendering, along with sampling weights.