Analyze, Generate, Improve: Failure-Based Data Generation for Large Multimodal Models

Gabriela Ben-Melech Stan[♡] Estelle Aflalo[♡] Avinash Madasu[♡] Vasudev Lal[♡] Phillip Howard[♢†]
[♡]Intel Labs [♢]Thoughtworks
{gabriela.ben.melech.stan,estelle.aflalo}@intel.com phillip.howard@thoughtworks.com

Abstract

Training models on synthetic data is an effective strategy for improving large multimodal models (LMMs) due to the scarcity of high-quality paired image-text data. Existing methods generate multimodal datasets but do not address specific reasoning deficiencies in LMMs. In contrast, humans learn efficiently by focusing on past failures. Inspired by this, we propose a synthetic data generation approach that analyzes an LMM's reasoning failures using frontier models to generate and filter high-quality examples. Our method produces a 553k-example multimodal instruction tuning dataset, leading to improved LMM performance, even surpassing models trained on equivalent real data demonstrating the high value of generating synthetic data targeted to specific reasoning failure modes in LMMs.

1. Introduction

Recent advancements in large language models (LLMs) can be attributed largely to scaling models and training data. However, high-quality data availability limits further scaling [38, 41]. As a result, synthetic data generation is gaining traction for augmenting datasets. High-quality synthetic data generation has grown more feasible with increasing LLM capabilities, enabling synthetic data production at scale [7, 13, 14, 20, 34, 44]. Synthetic data is particularly valuable for training large multimodal models (LMMs), which combine an LLM with a vision encoder to enable text generation conditioned on multimodal inputs. Real data for training LMMs is relatively scarce due to the lack of naturally-occurring images paired with high-quality text.

Existing approaches for generating multimodal synthetic data face two key limitations. First, they depend on real images paired with synthetic text generated by another LMM, restricting their use when image data is scarce. Second, they generate data arbitrarily without prioritizing useful examples, leading to inefficiencies in both data generation and training. For a broader discussion of previous methods, see Section 6 in the supplementary material. Unlike standard synthetic data generation, humans learn efficiently by focusing on examples tied to past reasoning failures. Ericsson et al. [8] suggest expert learning benefits from explicit instruction, error diagnosis, and targeted feedback. Since humans learn from failures [3, 6], seeking examples related to past errors aids mastery, while problems involving familiar reasoning are often ignored.

Inspired by this observation, we propose a synthetic data generation approach based on an LMM's reasoning failures. We first evaluate the LMM on a benchmark dataset and use a strong frontier model to analyze its errors. The frontier model then generates related question-answer pairs and image descriptions, which can link to existing images or guide synthetic image generation. Finally, we ensure quality using an LMM-as-a-judge filtering process. Our approach generates a 553k-example multimodal dataset from LLaVA-1.5-7B's [22] reasoning failures. Training experiments show our synthetic data improves LLaVA's performance across various downstream tasks, even surpassing training on real datasets, unlike prior work requiring significantly more examples to match real data performance [12]. Our method achieves greater gains compared to previously-proposed synthetic datasets, demonstrating the value of grounding synthetic data in model failure analysis.

2. Dataset construction

Diagnosing model failures To generate our tailored synthetic data, we first analyze reasoning failures in a baseline LMM using a more advanced frontier LMM. The frontier model is selected for its superior multimodal reasoning capabilities and high accuracy on diverse vision-language benchmarks. Reasoning failures are identified by evaluating both models on the training sets of vision-language benchmarks, and selecting samples where the baseline LMM produces incorrect responses, while the frontier model succeeds. This process generates a subset of failure cases per benchmark, providing a focused challenging dataset, which

[†]Work completed while at Intel Labs.

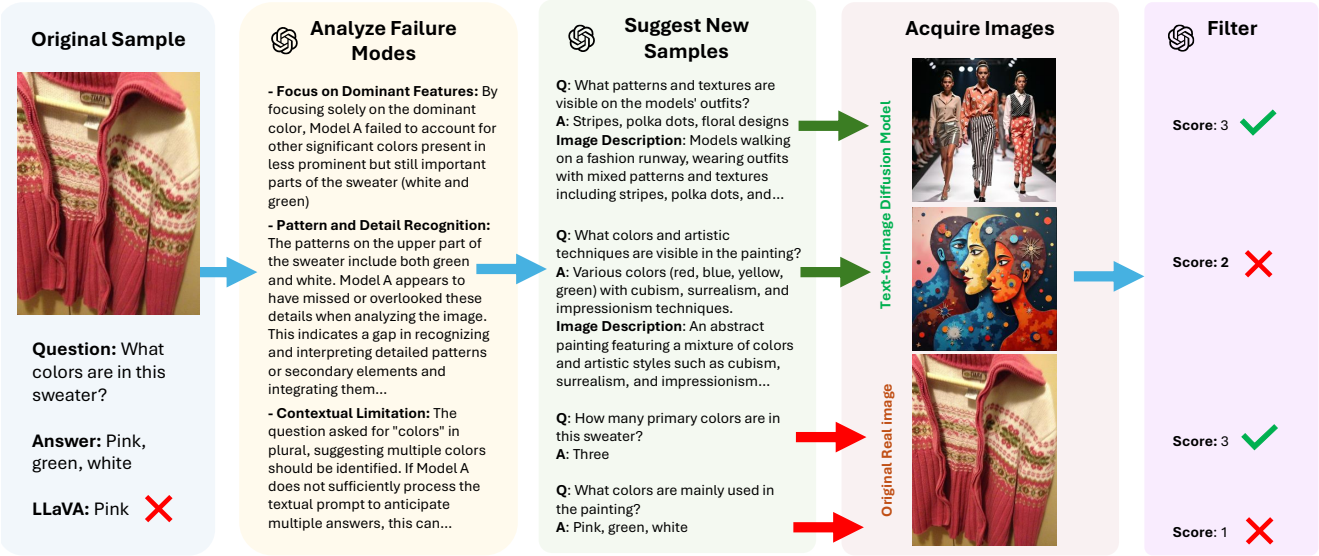


Figure 1. Illustration of our approach. Given a sample from an existing dataset which LLaVA answers incorrectly, we prompt a frontier model to analyze LLaVA's reasoning failures and propose new synthetic samples which require similar types of reasoning.

we denote as Model Failure Sets (MFS).

Synthetic data generation Using the resulting MFS, we guide the frontier model through a structured multi-turn process to diagnose the failure reasons of the baseline LMM and generate new training samples that address these failure modes, as illustrated in Figure 2 in Section 7.3 of the supplementary material (SM). In the first steps, the frontier model is prompted to describe the image and analyze reasoning errors by examining the question, ground truth answer, and the incorrect response generated by the baseline LMM. Next, the frontier model is instructed to propose new challenging samples designed to target the identified failure modes, which consist of a detailed image description, a clear question, and a deterministic answer. We explore two different approaches for sourcing images: utilizing existing real images (Method 1), or using synthetically generated images (Method 2).

Specifically, in Method 1, we leverage the original image from a failed sample and prompt the frontier model to generate 10 new question-answer pairs per sample, following the prompt in Figure 2, omitting the final step. This is especially effective for benchmarks like InfoVQA and ScienceQA, since text-to-image models often struggle with precise text rendering and spatial accuracy. In Method 2, the frontier model generates a question-answer pair and a detailed image prompt using the prompt in Figure 2. Each image prompt is then fed into a text-to-image diffusion model to generate 10 synthetic images at varying guidance scales, producing 100 fully synthetic samples per failed sample.

To enhance data diversity, we use a variation of our prompt instructing the frontier LMM to "provide examples that challenge Model A's weaknesses using scenarios from entirely different domains or situations". This instruction is added to Step 4 in Figure 2 to encourage samples generation in different domains. Domain-similar samples preserve the original theme, while non-similar samples offer broader contextual diversity for improved generalization, see Figure 6 in Section 7.3 (SM). We also enforce constraints on question format and instructions, as detailed in Section 7.3.

Filtering We apply a filtering process using the same frontier LMM which produced the samples. The LMM is instructed to evaluate each synthetic sample given an image, question, and answer on a scale of 1 to 3, where 1 indicates an incorrect sample, 2 being partially correct, and 3 fully correct. The prompt used for filtering is provided in Figure 3 in Section 7.4 (SM). Only samples rated 3 were included in the final dataset to ensure quality and reliability.

Dataset overview Our synthetic dataset consists of 553,992 samples incorporating both real and generated images derived from the MFS of LLaVA-1.5-7B, with Vicuna-1.5-7B [46] base LLM, on four benchmark training sets: VizWiz [11], InfoVQA [27], ScienceQA [25], and OK-VQA [26]. These benchmarks were chosen to ensure a wide range of visual and reasoning challenges, as detailed in Section 7.2 (SM). We use GPT-4o [30] as the frontier LMM for analyzing the reasoning failures of LLaVA-7B due to its strong multimodal reasoning capabilities, and FLUX.1-schnell Labs [18] text-to-image model to generate the synthetic images. Table 3 in Section 7.2 (SM) provides an

Dataset	Model	N	N_{syn}	EM Score
InfoVQA	LLaVA	624,610	0	26.7
	$\text{LLaVA}_{\text{real}}$	634,684	0	31.6
	$\text{LLaVA}_{\text{syn}}$	634,684	10,074	30.8
	$\text{LLaVA}_{\text{syn}}$	687,071	62,461	33.0
	$\text{LLaVA}_{\text{syn}}$	710,610	86,000	**34.3**
ScienceQA	LLaVA	624,610	0	70.7
	$\text{LLaVA}_{\text{real}}$	630,195	0	70.0
	$\text{LLaVA}_{\text{syn}}$	630,195	5,585	71.8
	$\text{LLaVA}_{\text{syn}}$	646,594	21,984	**73.0**
OK-VQA	LLaVA	624,610	0	57.0
	$\text{LLaVA}_{\text{real}}$	633,619	0	54.3
	$\text{LLaVA}_{\text{syn}}$	633,619	9,009	61.3
	$\text{LLaVA}_{\text{syn}}$	687,071	62,461	61.3
	$\text{LLaVA}_{\text{syn}}$	749,532	124,922	**61.5**

Table 1. In-domain evaluation of baseline LLaVA, LLaVA models trained using synthetically augmented data ($\text{LLaVA}_{\text{syn}}$) and using training data augmented with real in-domain data ($\text{LLaVA}_{\text{real}}$).

overview of each benchmark, including the original size of the training split, the MFS size of LLaVA-7b, and the total number of synthetic samples retained after filtering.

As previously described, we utilize two methods for synthetic data generation, depending on the dataset: for InfoVQA and ScienceQA, we apply Method 1, which generates new question-answer pairs based on the original real images. For OK-VQA, we employ Method 2, generating synthetic images along with corresponding question-answer pairs, and for VizWiz we apply both methods. Overall, our final dataset consists of 42% real-image-based samples and 58% fully synthetic samples. All samples in the final dataset contain a single-turn conversation, featuring a question and a short answer. Our filtering approach effectively removes lower-quality synthetic samples, with synthetic-image samples exhibiting higher removal rates compared to real-image samples, further details in Section 7.2 (SM). Examples of our dataset can be seen in Section 11 (SM).

Additionally, the large volume of filtered data ensures broad coverage of LLaVA-7B's reasoning failures across benchmarks, as listed in Section 9.5 (SM).

3. Experiments

Details of experimental setting We fine-tune LLaVA-1.5-7B using a subset of the LLaVA-Instruct-1.5-mix-665K dataset [23], which contains 665K user-GPT conversations focused on visual prompts. Since our approach targets image-grounded reasoning, we use only the 624K samples with images. We then augment this visually oriented subset with our synthetic MFS data (Section 2). The final training set combines real image-text conversations with targeted synthetic examples to address reasoning failures.

For fine-tuning, we used Vicuna-1.5-7B weights as the LLM backbone and leveraged the pretrained multimodal projector from LLaVA-1.5-7b. We followed the original LLaVA training procedure, ensuring a fair comparison to existing methods, with training details provided in Section 8 (SM). We refer to models trained with a mixture of the original LLaVA-Instruct dataset and our synthetic data as $\text{LLaVA}_{\text{syn}}$. As a baseline, we report the performance of LLaVA trained under the same setting but without any additional synthetic data added to the training dataset (i.e., $N_{syn} = 0$). Additionally, for each dataset from which reasoning failures were derived for synthetic data generation, we report the performance of a LLaVA model trained on an equivalent amount of real data sourced from the corresponding training dataset (denoted as $\text{LLaVA}_{\text{real}}$). This provides a measure of the efficiency of our synthetic data relative to training on real in-domain data.

We used a variety of multimodal reasoning benchmarks for evaluating models. For in-domain evaluations (i.e., evaluating on a withheld validation set corresponding to the training set from which reasoning failures were derived), we used the InfoVQA, OK-VQA and ScienceQA validation sets. Since InfoVQA and ScienceQA questions rely heavily on reading text contained in the images, we further evaluated models on TextVQA [33] and OCR-Bench [9]. Finally, because our synthetic dataset was designed to enhance the model's reasoning capabilities, we chose MMBench [24] and MMMU [45] as additional OOD benchmarks.

Synthetic data augmentation results Table 1 provides in-domain evaluation results utilizing synthetic data derived from InfoVQA, ScienceQA, and OK-VQA. Notably, augmenting the LLaVA-Instruct dataset with our synthetic data achieves performance comparable to or better than using an equivalent amount of real domain-specific data in most cases. This result is particularly significant given that the synthetic samples were generated using only a small subset of the original training data: specifically, only those examples where LLaVA scored 0.0 while GPT scored 1.0. For instance, with OK-VQA, the original training set consists of 9,009 samples, but we utilized only 607 training samples which LLaVA failed on in order to generate 9009 synthetic samples, resulting in a performance boost of 13% on the OK-VQA test set. Similarly, our approach utilized only 28% of the ScienceQA training dataset to generate full synthetic replacements, yet still resulted in better performance than training directly on the real dataset. We also observe that performance improves as the amount of synthetic data used for data augmentation increases. In practice, it may be desirable to combine synthetically generated data which was derived from reasoning failures across different datasets. We therefore provide results for two different sized mixtures of our synthetic data derived from InfoVQA, ScienceQA, and OK-VQA reasoning failures in Table 2. In addition to in-domain evaluations for these three datasets, we provide results for the four other datasets mentioned

Base LLM	N	N_{syn}	Augmentation Data	TextVQA	OCR-Bench	InfoVQA	OK-VQA	ScienceQA	MMBench	MMMU
Vicuna-7B	624,610	0	N/A (baseline)	47.0	31.9	26.7	57.0	70.7	52.3	36.4
Vicuna-7B	687,071	62,461	ALLaVA	**47.9**	**34.0**	28.4	50.4	71.2	50.2	**36.7**
			CoSyn-400K	47.1	31.8	28.5	55.8	71.5	52.4	34.6
			SimVQA	46.8	31.6	26.6	54.4	71.1	**53.5**	34.7
			Ours	47.4	33.2	**33.1**	**60.8**	**73.1**	52.5	36.2
Vicuna-7B	749,532	124,922	ALLaVA	47.2	34.1	28.8	49.4	66.5	43.5	34.2
			CoSyn-400K	46.8	32.7	29.6	57.7	70.9	51.9	36.9
			Ours	**47.4**	**34.5**	**33.2**	**61.1**	**73.0**	**52.5**	**37.4**
Gemma-2B	624,610	0	Ours	39.9	28.3	21.8	51.7	62.3	29.2	**32.3**
	749,532	124,922		**40.9**	**29.9**	**29.8**	**54.8**	**65.2**	**30.7**	31.7
Qwen2-7B	624,610	0	Ours	45	31.4	26.7	59.2	77.9	63	**42.3**
	749,532	124,922		**46.2**	**32.5**	**27**	**60.6**	**80.7**	**63.5**	42.2

Table 2. Training data augmentation experimental results. N denotes the total number of training examples, N_{syn} denotes the number of synthetic examples in the training dataset generated using our approach. The first section of the table compared the same base LLM (Vicuna) trained on various datasets with our dataset, while the second section compares different LLM backbones trained on our dataset.

previously to measure the impact on OOD generalization. We also provide results for LLaVA models trained on three alternative synthetically generated datasets: ALLaVA, CoSyn-400k, and SimVQA. From Table 2, we observe that our synthetic data outperforms all baselines in the maximum data augmentation setting ($N_{syn} = 124,922$). When augmenting training sets with half as much data, we observe that our synthetic data produces the greatest improvements in-domain, whereas the ALLaVA dataset, which contains only real images paired with synthetic text, performs slightly better in other OOD settings. Finally, training LLaVA on the mixed synthetic dataset yields similar in-domain performance to training it separately on failures from each individual dataset (Table 1).

To evaluate the generalization ability of our synthetic dataset, we trained models using the same datasets but with different backbone LLMs. In the following experiments, we used Gemma-2B [35] and Qwen2-7B [42] as base LLMs. We adopted the LLaVA two-phase training procedure: pretraining on the LLaVA 558k dataset [23] followed by instruction fine-tuning. Table 2 indicates that our synthetic data in its maximum augmentation setting, $N_{syn} = 124,922$, outperform the baseline for both LLaVA-Gemma-2B and LLaVA-Qwen-7B almost on all benchmarks. Notably, despite being generated based on LLaVA-Vicuna-7B failure mechanisms, our synthetic data enhance the performance of other models whether they are of same size (Qwen2-7B) or smaller (Gemma-2B).

4. Analysis and Ablations

We analyzed the quality of our dataset, as detailed in Section 9 (SM). We conducted a human evaluation of our dataset (Section 9.1) and found that fully synthetic samples match or slightly exceed the fidelity of real-image-based sample. The evaluation also demonstrates excellent alignment between generated questions, answers, and image prompts. In Section 9.2, we tested our data in low resource setting by replacing portions of the original dataset with fully synthetic samples. Even with up to 25% substitution, the model finetuned on our dataset matches or outperforms the baseline LLaVA, highlighting the efficiency of our targeted dataset. Section 9.3 quantitatively describes the impact of filtering on data quality and shows that our filtering approach improves model performance. Section 9.4 compares two frontier LMMs for synthetic data generation, showing that Qwen2-VL [39] produces samples with lower downstream performance than GPT-4o (See example in Figure 5). Finally, Section 9.5 identifies reasoning failures and shows that targeted synthetic data improves LLaVA-Instruct's performance on related benchmarks.

5. Conclusion

We introduced a new approach for generating multimodal synthetic data based on analyzing a model's reasoning failures. This led to a multimodal instruction tuning dataset with over 553k synthetic examples derived from LLaVA's failures. Experiments show that our data significantly improves LLaVA's performance on InfoVQA, ScienceQA, and OK-VQA, even surpassing training on an equivalent amount of real data. Furthermore, models trained on our synthetic dataset exhibit improvements in OOD evaluations and outperform training on other existing synthetic datasets when training data augmentation is scaled, showing consistent improvements across different base models. We also showed that training LLaVA only on examples derived from specific failure modes improves its performance on tasks which require corresponding forms of reasoning. Ablations and human evaluations confirm the effectiveness and quality of our approach, highlighting the potential of targeted synthetic data generation to address model deficiencies.

References

[1] Deepseed AI. Deepspeed. https://github.com/deepspeedai/DeepSpeed, 2025. Accessed: 2025-03-07. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 4

[3] John Seely Brown, Allan Collins, and Paul Duguid. Situated cognition and the culture of learning. *1989*, 18(1):32–42, 1989. 1

[4] Atoosa Chegini and Soheil Feizi. Identifying and mitigating model failures through few-shot clip-aided diffusion generation. *arXiv preprint arXiv:2312.05464*, 2023. 1

[5] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024. 1

[6] Aubteen Darabi, Thomas Logan Arrington, and Erkan Sayilir. Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development*, 66:1101–1118, 2018. 1

[7] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Luu Anh Tuan, and Shafiq Joty. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705, 2024. 1

[8] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363, 1993. 1

[9] Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2024. 3

[10] Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. Targen: Targeted data generation with large language models. *arXiv preprint arXiv:2310.17876*, 2023. 1

[11] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 2, 1

[12] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 1, 3

[13] Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. *arXiv preprint arXiv:2210.12365*, 2022. 1

[14] Phillip Howard, Junlin Wang, Vasudev Lal, Gadi Singer, Yejin Choi, and Swabha Swayamdipta. Neurocomparatives: Neuro-symbolic distillation of comparative knowledge. *arXiv preprint arXiv:2305.04978*, 2023. 1

[15] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022. 1

[16] Gyeongman Kim, Doohyuk Jang, and Eunho Yang. Promptkd: Distilling student-friendly knowledge for generative language models via prompt tuning. *arXiv preprint arXiv:2402.12842*, 2024. 1

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[18] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 1

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4

[20] Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*, 2024. 1

[21] Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. Synthesize step-by-step: Tools templates and llms as data generators for reasoning-based chart vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13613–13623, 2024. 1

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 3, 4

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 3

[25] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2, 1

[26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 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, 2019. 2, 1

[27] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa.

In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2, 1

[28] Rakesh R Menon and Shashank Srivastava. Discern: Decoding systematic errors in natural language for text classifiers. *arXiv preprint arXiv:2410.22239*, 2024. 1

[29] Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Weige Chen, Olga Vrousgos, Corby Rosset, et al. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*, 2024. 1

[30] OpenAI. Gpt-4o system card, 2024. 2

[31] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 4

[32] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 3

[33] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3

[34] Xin Su, Man Luo, Kris W Pan, Tien Pei Chou, Vasudev Lal, and Phillip Howard. Sk-vqa: Synthetic knowledge generation at scale for training context-augmented multimodal llms. *arXiv preprint arXiv:2406.19593*, 2024. 1

[35] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. 4

[36] Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*, 2023. 1

[37] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 4

[38] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 1, 2022. 1

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 4, 3

[40] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 4

[41] Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36:59304–59322, 2023. 1

[42] 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, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. 4

[43] Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, et al. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. *arXiv preprint arXiv:2502.14846*, 2025. 1

[44] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator:

A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36:55734–55784, 2023. 1

[45] 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, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. 3

[46] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2

[47] Ingo Ziegler, Abdullatif Köksal, Desmond Elliott, and Hinrich Schütze. Craft your dataset: Task-specific synthetic dataset generation through corpus retrieval and augmentation. *arXiv preprint arXiv:2409.02098*, 2024. 1

Analyze, Generate, Improve: Failure-Based Data Generation for Large Multimodal Models

Supplementary Material

6. Related Work

Synthetic datasets for training LMMs. Chen et al. [5] introduced ALLaVA, a 1.3M-sample dataset of real images with annotations and QA pairs from a frontier LMM, but it lacks failure-driven data generation and synthetic images. Li et al. [21] generate synthetic QA pairs for real chart images, focusing on chart VQA. Yang et al. [43] use code-guided generation (e.g., LaTeX, HTML) to create text-rich synthetic images. In contrast, our approach is broadly applicable across domains and leverages text-to-image diffusion models for greater image diversity.

Synthetic data generation from model failures. Prior work has explored leveraging model failures for synthetic data generation. Jain et al. [15] identify failure-related directions in a vision model's latent space to guide diffusion models in generating corrective images. Chegini and Feizi [4] use ChatGPT and CLIP to generate text prompts for diffusion models based on vision model failures. In language models, DISCERN [28] iteratively describes errors for synthetic data generation, while Lee et al. [20] uses incorrect answers from a student LLM finetuned on specific tasks as input to a teacher LLM which generates new examples to use for training. Unlike prior work, which generates single-modality data (text-only or image-only) and focuses on classification tasks, our approach generates multimodal image-text datasets aimed at training models for open-ended text generation.

Generating synthetic data from frontier models to teach new skills. AgentInstruct [29] is an agentic framework for generating synthetic data from a powerful frontier model (e.g., GPT-4) to teach new skills to a weaker LLM. Similarly, Ziegler et al. [47] utilize few-shot examples annotated by humans and retrieved documents with produce synthetic data from LLMs for teaching specialized tasks to models. Prompt-based methods for synthetic data generation from LLMs without seed documents [10, 36] as well as knowledge distillation from a teacher model [16] have also been proposed. Unlike our work, these prior studies focus on language-only data generation and use seed documents (e.g., raw text, source code) or prompts as a basis for data generation rather than an analysis of model failures.

7. Dataset generation

7.1. Compute Infrastructure

To generate our dataset, we queried GPT-4o through the Azure OpenAI API and deployed Qwen2-VL on Nvidia RTX A6000 GPUs. Using Intel® Gaudi 2 AI accelerators from the Intel® Tiber™ AI Cloud, we generated 1.024 million images from the VizWiz failed samples and 535k images derived from OK-VQA.

7.2. Dataset statistics

Our synthetic dataset is derived from the MFS of LLaVA-1.5-7B on four benchmark training sets: VizWiz [11], InfoVQA [27], ScienceQA [25], and OK-VQA [26], selected to cover diverse visual and reasoning challenges. VizWiz consists of real-world images captured by visually impaired users, often requiring detailed scene understanding. OK-VQA focuses on visual question answering which requires external knowledge. InfoVQA involves text-rich images where reading comprehension is crucial, assessing the model's ability to extract and interpret textual information from images. ScienceQA includes multimodal scientific reasoning questions which require both spatial and logical reasoning, making it valuable for evaluating complex reasoning capabilities. To generate the synthetic images, we utilized FLUX.1-schnell Labs [18] text-to-image model with a resolution of 1024×1024 pixels and guidance scale range of 3 to 13.

Table 3 provides statistics detailing the quantity of synthetic examples in our dataset which were derived from reasoning failures on different benchmarks. Additional discussion of the dataset composition is provided in Section 2.

Our filtering approach successfully removes poor-quality samples, with the following removal rates across benchmarks: for VizWiz, 81% of synthetic-image samples and 34% of real-image samples were removed. This indicates that generating entirely synthetic samples is more challenging than generating synthetic text alone for real images. OK-VQA had a lower removal rate of 29% for synthetic images, possibly resulting from simpler and less ambiguous visual content. Among real-image-based samples, ScienceQA experienced a similar removal rate (29%), likely due to the complexity of spatial and scientific reasoning tasks. In contrast, InfoVQA exhibited a significantly lower removal rate of only 5% with an average filtering score of 2.9 (out of 3), indicating the strong capability of GPT-4o in handling text-based images.

$\text{Dataset}_{\text{Image Type}}$	Original	Failures	Filtered
$\text{VizWiz}_{\text{real}}$	20,523	7,785	100,280
$\text{VizWiz}_{\text{syn}}$	20,523	7,785	190,172
$\text{InfoVQA}_{\text{real}}$	10,074	5,250	95,783
$\text{ScienceQA}_{\text{real}}$	5,585	1,562	39,090
$\text{OK-VQA}_{\text{syn}}$	9,009	607	128,667

Table 3. Dataset statistics across benchmarks, including original training set size, number of failure samples (LLaVA-1.5-7b: 0, GPT-4o: 1), and synthetic samples with filtering score 3.

7.3. Data generation prompt

Figure 2 shows the prompt used to generate fully synthetic question-answer, image samples based on the failure modes of an LMM. To enhance data diversity, we use a variation of our prompt, expending step 4 to generate examples in different domains. Figure 6 compares fully synthetic samples with generated images, within similar and non-similar domain of the original failed sample. we notice that domain-similar samples preserve the original theme, while non-similar samples cover a more diverse contextual range to improve generalization. Additionally, we created samples where we both enforced and relaxed constraints on question format (e.g., multiple-choice, true/false) and instructions (e.g., requiring responses like "Unanswerable" when information was insufficient or limiting answers to short responses, see the Shiba Inu example from Figure 9).

7.4. Filtering prompt

Figure 3 provides the prompt which we used for the filtering stage of our synthetic data generation pipeline. See Section 2 of the main paper for additional filtering details.

8. Training hyperparameters

To train our model, we used 8 Nvidia RTX A6000 GPUs using the hyperparameters from Table 4. We employed DeepSpeed ZeRO stage 3 [1] for distributed training.

Batch Size/GPU	16
Number of GPUs	8
Gradient Accumulation	1
Number of epochs	1
LLaVA Image Size	576
Optimizer	AdamW
Learning Rate	$2e-5$
BF16	$True$
LR scheduler	$cosine$
Vision Tower	openai/clip-vit-large-patch14-336
Language Model	lmsys/vicuna-7b-v1.5

Table 4. Hyperparameters to train our model.

```
You are analyzing the performance of
a vision-language model (called Model
A). Model A's answer could deviate from
the ground truth due to limitations in
visual understanding, interpretation, or
reasoning.
Step 1:  Describe the image.
Step 2:  Given a question, the Ground
truth answer, and Model A's generated
answer, describe any key visual
elements that might influence Model A's
interpretation.
Step 3:  Analyze the reasoning steps Model
A might have used to generate its answer,
considering both the visual and textual
information.  Identify any weaknesses,
errors, or gaps in Model A response
compared to the ground truth.
Step 4:  Suggest 10 additional challenging
detailed examples to address these
limitations.
Step 5:  Transform each example into a
detailed prompt designed to generate
a clear and realistic image using a
text-to-image generation model.
```

Figure 2. Prompt used to generate fully synthetic image-text samples based on the failure modes of an LMM (Method 2).

```
Given sample containing an image, a
question, and an answer, your task is
to grade the sample from 1 to 3 based
on the following criteria:
Score 1:  The answer is incorrect.
Score 2:  The answer is correct, but
it is one of several possible valid
answers.
Score 3:  The answer is correct,
specific, and the only valid answer.
The image provides all the necessary
context for the answer.
```

Figure 3. Filtering prompt

9. Additional Analysis

9.1. Human evaluation of dataset quality

Three of the authors of this work conducted a human evaluation by assessing three different aspects of our generated samples: (1) the alignment of the question and answer in relation to the image prompt, (2) the alignment between the image prompt and the generated image, and (3) the correctness of the answer given the question and image. The first evaluation reflects the quality of reasoning, the second evaluates the fidelity of the image generator's output, and the third combines both aspects. Scores range from 1 to

3, where 1 indicates an irrelevant alignment, 3 signifies a relevant alignment, and 2 represents a partially relevant or ambiguous alignment. We evaluated 200 samples in total, with 101 containing real images and 99 being fully synthetic. The overall correctness score for answers was 2.78, with real-image-based samples scoring 2.75 and fully synthetic samples scoring 2.81, indicating that fully synthetic samples achieve a level of fidelity equal to or even slightly exceeding that of real-image-based samples. For the synthetic samples specifically, we also measured the alignment between the image prompt and the generated image (2.66), and the alignment of the generated question and answer with the image prompt (2.84), indicating the high quality of reasoning in the generated responses.

9.2. Training data substitution vs. augmentation and impact of synthetically generated images

Our previous experiments augmented an existing 624k sample training dataset (LLaVA-Instruct) with our synthetic data. In domains where data is scarce, training datasets of this size may not be available. To investigate the utility of our synthetic data in such low-resource settings, we conducted experiments in which we randomly substituted different quantities of examples from the original dataset with our synthetically generated data*. The results of this experiment are provided in rows 2-3 of Table 5. Even when up to 25% of the original dataset is substituted with our synthetic data, we achieve performance that is either as good or better than the baseline LLaVA model across a broad range of downstream tasks. This is despite the fact that the original LLaVA training dataset utilizes real images, whereas our synthetic data used in this experiment contained only synthetically generated images. The fact that our synthetic data achieves similar or better performance than an existing real data source is significant, as prior studies have shown that training on synthetically generated image data is often much less efficient than training on an equivalent amount of real image data [12]. Table 5 also shows the impact of using real vs. synthetic images in our pipeline. Specifically, we compare the effectiveness of our synthetic data derived from Vizwiz reasoning failures when paired with real images (from Vizwiz) or synthetically generated images. In the training data augmentation setting, we observe that synthetic images generally achieve similar results as utilizing real images. Synthetic images even surpass the performance of real images in TextVQA, OK-VQA, and MMBench. This demonstrates the high quality of our synthetic images and their potential to serve as replacements for real images in low-resource settings where data is scarce.

*We used synthetic data derived from Vizwiz failures in this setting.

9.3. Impact of filtering on data quality

To investigate the impact of filtering on the quality of synthetically generated data, we repeated our in-domain evaluation experiments for ScienceQA and OK-VQA using raw unfiltered data. In the maximum synthetic data augmentation setting (last row of each section in Table 1), using unfiltered data reduces EM from 73.0 to 72.2 on ScienceQA and from 63.3 to 58.8 on OK-VQA. This shows that our filtering approach improves model performance when using our synthetic examples for training data augmentation. Furthermore, using only synthetic examples which were assigned the lowest rating in our filtering process decreases the EM score on OK-VQA to 57.5, which highlights the difference in quality between the lowest-scoring and highest-scoring synthetic examples identified during filtering.

9.4. Comparison of LLM synthetic data generators

We compared two frontier LMMs, GPT-4o and Qwen2-VL-7B [39], for generating synthetic data grounded in LLaVA-7B failures. Qwen2-VL-7B was selected due to its high accuracy on vision-language benchmarks. Our results show that using samples generated by Qwen2-VL leads to reduced downstream performance compared to those produced by GPT-4o, with a decrease of 2% on InfoVQA and 6.5% on OK-VQA. Additionally, samples generated by Qwen2-VL received lower filtering scores: 1.9 (Qwen2-VL) vs. 2.5 (GPT-4o) for OK-VQA, and 2.6 (Qwen2-VL) vs. 2.9 (GPT-4o) for InfoVQA. Based on our manual analysis, we hypothesiize that these differences may result from the detailed and precise reasoning provided by GPT-4o, resulting in synthetic samples that are better tailored to address identified reasoning failures. In contrast, samples generated by Qwen2-VL-7B sometimes demonstrate lower diversity, which could limit their effectiveness in addressing the broad range of failure modes. Figure 5 provides an example of these observed differences in the reasoning processes of GPT-4o and Qwen2-VL-7B models, as well as the corresponding generated fully synthetic samples.

9.5. Correcting specific types of reasoning failures

Our synthetic data generation approach explicitly identifies different types of LMM reasoning failures. To systematically categorize these failures, we encoded each reasoning explanation using sentence transformers [32] and clustered them using k-means. Figure 4 presents the resulting clusters, highlighting prevalent failure modes such as optical character recognition (OCR) and object detection errors. Based on this analysis, we further investigated whether targeted synthetic data can effectively address these specific failure cases and enhance LLAVA's reasoning capabilities.

Specifically, we augmented LLaVA-Instruct with 10,579 synthetic samples from our $\text{VizWiz}_{\text{syn}}$-MFS addressing object detection reasoning failures and repeated the second

Train Data	N	N_{syn}	TextVQA	OCR-Bench	InfoVQA	OK-VQA	ScienceQA	MMBench	MMMU
Baseline	624,610	0	47.0	31.9	26.7	57.0	70.7	52.3	36.4
Substitute w/	624,610	62,461	47.1	31.6	26.5	56.9	70.8	52.3	35.3
syn images	624,610	156153	46.9	31.1	27.0	57.0	70.6	51.2	**37.9**
Augment w/	645,222	20,612	46.7	32.0	27.0	57.4	**71.2**	**53.4**	34.9
syn images	687,071	62,461	**47.7**	31.8	25.8	**59.4**	**71.2**	52.3	36.4
Augment w/	645,222	20,612	46.9	**32.5**	**27.2**	57.4	70.6	53.1	35.0
real images	687,071	62,461	47.2	32.2	**27.2**	56.9	**71.2**	52.3	33.8

Table 5. Ablation experiments comparing baseline LLaVA to LLaVA models trained with synthetic data generated from VizWiz failures. We investigate substitution and augmentation strategies for synthetic data, as well as the use of synthetic vs. real images.

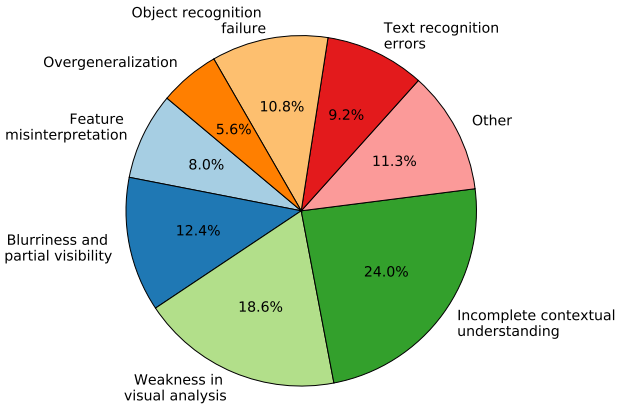


Figure 4. Figure shows the clusters of LLAVA reasoning failures described by GPT-4o.

Dataset	LLAVA	$LLaVA_{syn}$
CIFAR-10 [17]	**82.1**	81.2
Food-101 [2]	**13.4**	13.2
iNaturalist [37]	20.6	**52.0**
MNIST [19]	75.1	**80.5**
F-MNIST [40]	9.8	**10.0**
Oxford-pets [31]	39.6	**96.4**

Table 6. Image classification accuracy of LLaVA and a $LLaVA_{syn}$ model augmented only with synthetic examples corresponding to object recognition failures.

stage of LLaVA finetuning. The model was then evaluated on CIFAR-10 [17], Food-101 [2], iNaturalist [37], MNIST [19], Fashion-MNIST [40], and Oxford-Pets (Binary) [31] by formatting samples as multiple-choice questions. Table 6 presents a comparison of LLaVA and $LLaVA_{syn}$. The results show that $LLaVA_{syn}$ surpasses LLaVA on four out of six datasets, with particularly notable improvements on iNaturalist, MNIST and Oxford-Pets. This demonstrates the significant impact of our synthetic dataset in addressing specific reasoning failures within LLAVA. By systematically incorporating targeted synthetic samples, we can mitigate common failure cases, leading to measurable performance improvements across multiple benchmarks. Our findings highlight the effectiveness of leveraging targeted synthetic data to refine model reasoning and suggest that incorporating such data-driven interventions can significantly enhance the robustness and generalization of LMMs.

10. Detailed OOD results for models fit to different subsets of synthetically generated data

Table 7 provides additional evaluation results for models trained individually on real and synthetic data derived from Vizwiz, InfoVQA, ScienceQA, and OK-VQA. All reported values are the official evaluation metrics corresponding to each dataset. The first two rows of each section in Table 7 provide a direct comparison of the efficiency of our synthetic data to real data; we observe that augmenting the LLaVA-Instruct dataset with our synthetic data achieves as good or better performance across most settings as augmenting with real domain-specific data. Furthermore, significant performance gains are achieved relative to the LLaVA baseline when our synthetic data is derived from a dataset in the same domain as the benchmark. For example, synthetic data generated from reasoning failures on InfoVQA significantly improve LLaVA's performance on tasks which require fine-grained text understanding such as OCR-Bench and InfoVQA.

11. Examples from our dataset

In this section, we present examples from our dataset and highlight its weaknesses and limitations. Figure 6 shows a comparison of fully synthetic similar vs non-similar samples. Figures 7 show sampled examples from VizWiz and InfoVQA highlighting the diversity of question types and demonstrating the overall quality of generated images and text. Figure 9 shows our synthetic data generated from

Train Dataset	N	N_{syn}	TextVQA	OCR-Bench	InfoVQA	OK-VQA	ScienceQA	MMBench	MMMU
Baseline	624,610	0	0.47	0.32	0.27	0.57	0.71	52.30	0.36
Vizwiz	645,133	0	0.47	0.28	0.26	**0.59**	**0.71**	51.74	**0.38**
	687,071	62,461	**0.48**	**0.32**	0.26	**0.59**	**0.71**	52.25	0.36
	749,532	124,922	0.47	**0.32**	**0.27**	**0.59**	0.70	**53.02**	0.37
InfoVQA	634,684	0	0.47	0.32	0.32	0.58	0.70	52.50	0.37
	634,684	10,074	0.47	0.33	0.31	**0.59**	0.70	52.16	0.36
	687,071	62,461	0.47	**0.34**	0.33	0.57	**0.71**	**52.69**	0.37
	710,610	86,000	**0.48**	0.33	**0.34**	0.56	**0.71**	52.53	**0.38**
ScienceQA	630,195	0	0.47	0.29	0.26	**0.58**	0.70	52.88	0.36
	630,195	5,585	0.47	**0.32**	**0.27**	0.57	0.72	**53.19**	0.37
	646,594	21,984	0.47	**0.32**	0.26	0.56	**0.73**	53.12	**0.38**
OK-VQA	633,619	0	**0.47**	0.30	0.27	0.54	**0.71**	**53.35**	**0.36**
	633,619	9,009	**0.47**	**0.33**	**0.28**	**0.61**	**0.71**	52.68	0.35
	687,071	62,461	**0.47**	**0.33**	0.27	**0.61**	**0.71**	51.96	0.35

Table 7. Training data augmentation experimental results. N denotes the total number of examples used for training, while N_{syn} denotes the number of synthetic examples in the training dataset which were generated using our approach.

Figure 5. Comparison of GPT-4o and Qwen2-VL for generating Failure-Grounded Synthetic Datasets: GPT-4o demonstrates stronger reasoning capabilities, identifying multiple reasoning failures such as missed contextual cues and a lack of correlation between visual elements and the question. In contrast, while Qwen2-VL correctly answering the original question, identifies fewer failure modes and is less accurate in diagnosing LLaVA's reasoning failures, sometimes focusing on less relevant aspects, such as the kite in the sky. As a result, Qwen2-VL's generated samples are less diverse, often repeating the same question, whereas GPT-4o's samples provide broader coverage of identified reasoning failures. Note: GPT-4o's reasoning is 2–3 times longer than Qwen2-VL's; only a portion of GPT-4o's reasoning is shown here, while Qwen2-VL's reasoning is presented in full.

the OK-VQA dataset, while Figure 10 corresponds to the VizWiz dataset. These are fully synthetic examples, including generated image, along the question and answer. Figure 8 provides additional fully synthetic text & image examples derived from VizWiz and OK-VQA.

Figure 11, 12 and 13 illustrate examples derived from the ScienceQA, InfoVQA and VizWiz benchmarks respectively, where the images are real but the questions and answers are synthetically generated.

Lastly, Figures 14 and 15 show some incorrect examples for each benchmark.

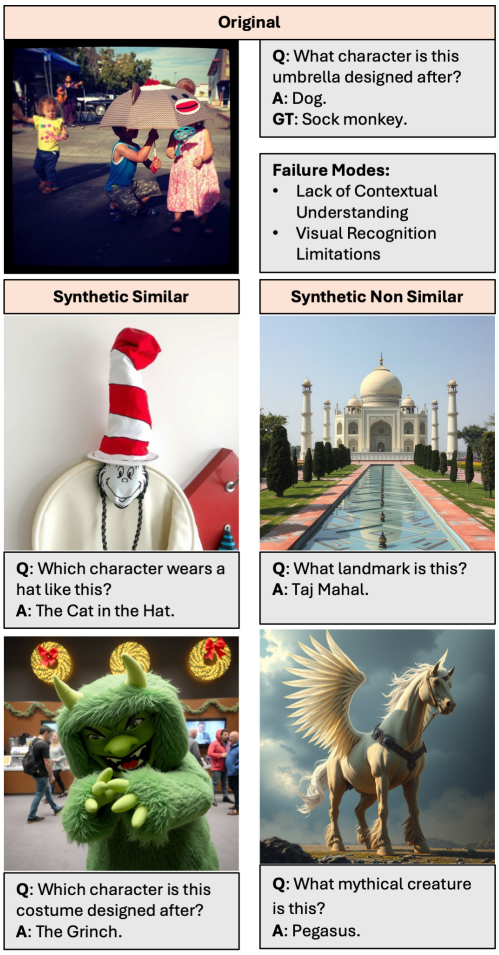


Figure 6. Comparison of fully synthetic similar and non-similar samples. Similar samples maintain a children's characters-based theme like the original sample, while non-similar samples address the failure modes by introducing diverse contexts.

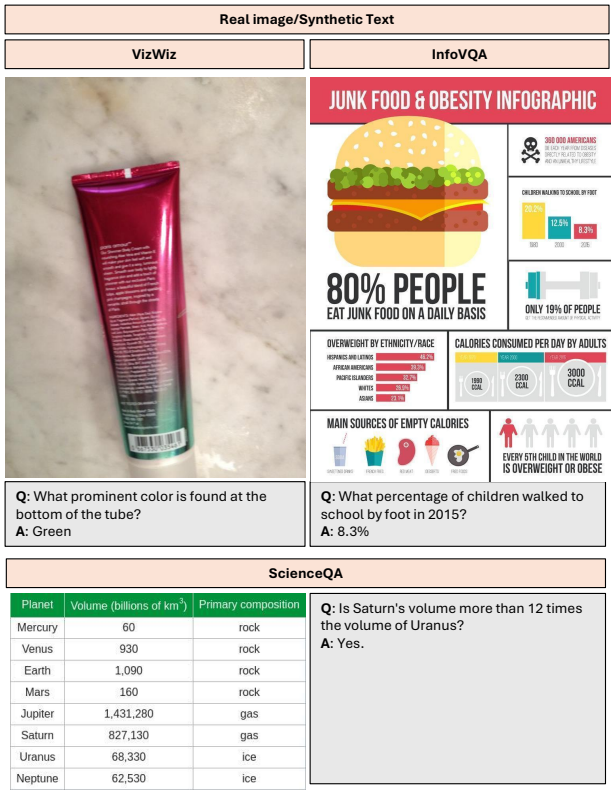


Figure 7. Examples of generated synthetic question-answer pairs for real images from VizWiz, InfoVQA, and ScienceQA.

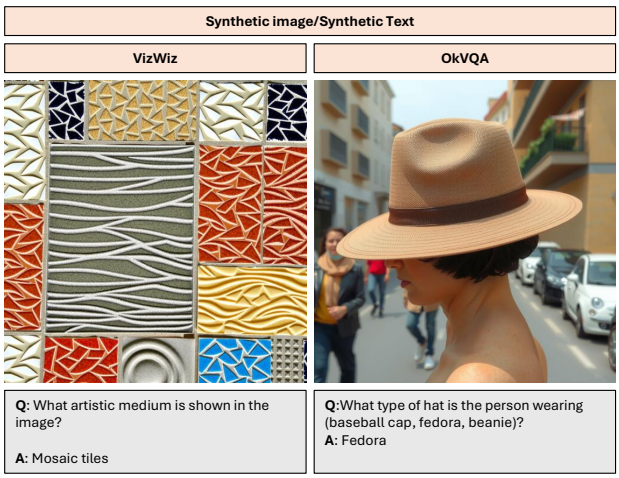
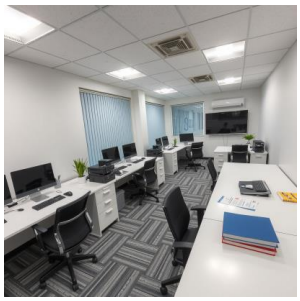


Figure 8. Examples of fully synthetic samples, using Method 2 as described in 2, both question-answer pairs and images were generated.

OkVQA– Synthetic Image/Synthetic Text

Q: What equipment is the person using to catch fish?
A: Fishing rod
Prompt: A person standing on a boat under a clear sky, casting a fishing rod into the water, with fishing gear and a cooler in the background

Q: Which breed is generally smaller and has a more fox-like appearance? When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase?
A: A Shiba Inu
Prompt: A Shiba Inu with a smaller, fox-like appearance and a curled tail, walking in a Japanese garden.

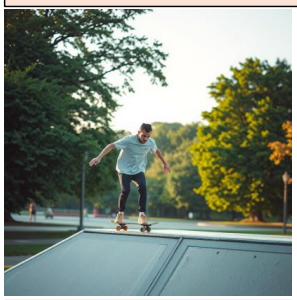


Q: Is it more likely to find a coffee mug or a pillow in this room?
A: Coffee mug
Prompt: A modern office space with desks, chairs, computers, and office supplies, arranged in a professional and organized manner, with some papers and stationery items on the desks.

Q: What type of clothing is this?
A: Kimono
Prompt: A traditional Japanese kimono displayed on a mannequin, with intricate designs and vibrant colors. The background is minimal, with a soft-focus effect. The kimono's patterns and texture are detailed and realistic

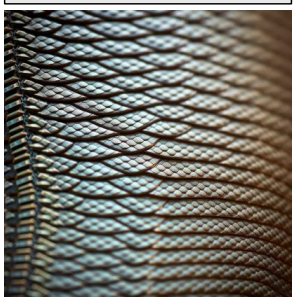
Figure 9. Examples of generated samples from OK-VQA with synthetic images and synthetic text.

VizWiz – Synthetic Image/Synthetic Text



Q: What sport is this individual engaged in?
A: Skateboarding
Prompt: A person skateboarding in a park, mid-action on a skateboard ramp. The background includes trees, a clear sky, and other park elements. The skateboarder is wearing casual streetwear and protective gear

Q: What musical instrument is this?
A: Violin
Prompt: A violin lying in its case with a bow next to it. The background is a wooden surface. The image is detailed, showing the strings, tuning pegs, and fine tuners of the violin.

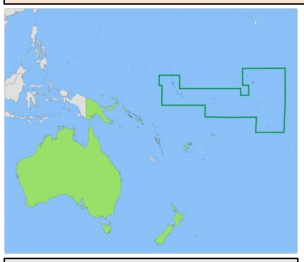


Q: What type of animal skin is shown in the image?
A: Reptile scales
Prompt: A close-up image of reptile scales, showing their overlapping structure and textured surface.

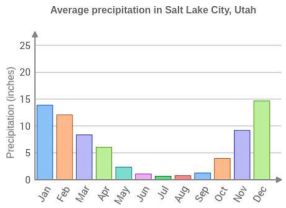
Q: What type of clothing is this?
A: Kimono
Prompt: A traditional Japanese kimono displayed on a mannequin, with intricate designs and vibrant colors. The background is minimal, with a soft-focus effect. The kimono's patterns and texture are detailed and realistic

Figure 10. Examples of generated samples from VizWiz with synthetic images and synthetic text.

ScienceQA– Real Image/Synthetic Text



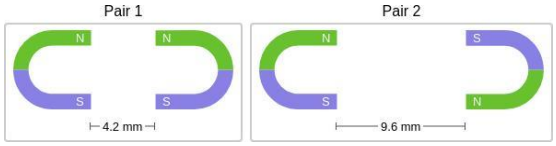
Q: Which country's maritime boundary is outlined in green?
A: Kiribati



Q: Which month has the highest average precipitation in Salt Lake City?
A: December

Aquarium	Initial temperature (°C)	Final temperature (°C)
Covered aquarium	21.7	16.2
Uncovered aquarium	21.7	16.2

Q: Based on the temperature data, did the aquariums lose or gain thermal energy?
A: Lose thermal energy

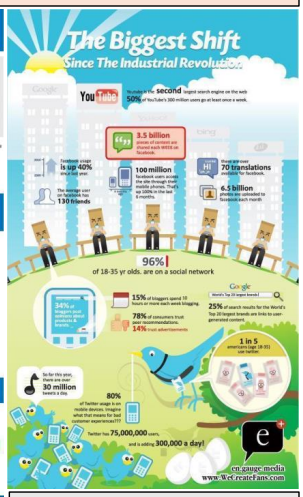


Q: In Pair 1, what color represents the North pole?
A: Green

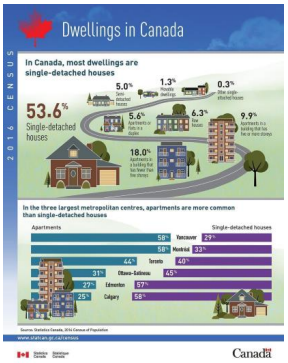
Figure 11. Examples of generated samples from ScienceQA with real images and synthetic text.

InfoVQA– Real Image/Synthetic Text

Q: What is the recommended distance to maintain from others after an earthquake?
A: 1 meter



Q: What percentage of bloggers spend 10 hours or more each week blogging?
A: 15%

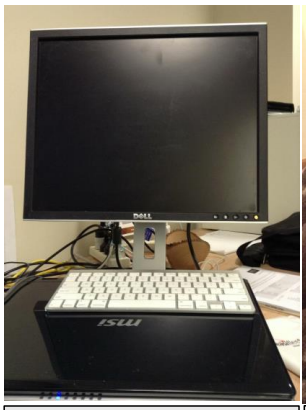


Q: What percentage of Canadian dwellings are row houses?
A: 6.3%

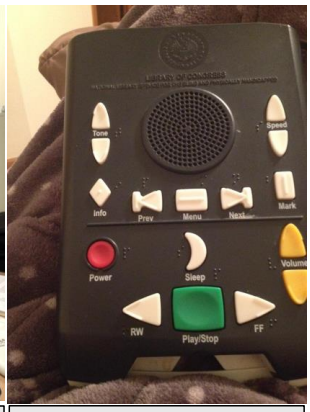
Q: What fraction of the grain grown in the US is used to feed farmed animals?
A: More than 70%

Figure 12. Examples of generated samples from InfoVQA with real images and synthetic text.

VizWiz– Real Image/Synthetic Text



Q: What brand is the monitor?
A: Dell



Q: How can a user navigate between different items?
A: Using Prev and Next buttons

Q: According to the package, is the "sweet caramel latte" artificially flavored?
A: Yes

Q: What category does this product belong to as indicated on the top left corner?
A: Tech

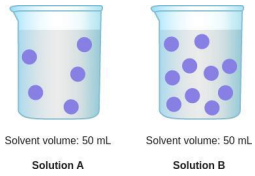
Figure 13. Examples of generated samples from VizWiz with real images and synthetic text.

Real image/Synthetic Text

VizWiz

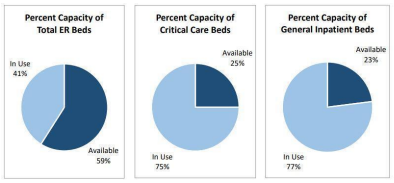
Q: What ingredient needs to be added to prepare the contents?
A: Water

ScienceQA



Q: By how many particles does Solution B exceed Solution A in terms of solute particles? A. 5 particles B. 7 particles C. 2 particles,
A: B.

InfoVQA

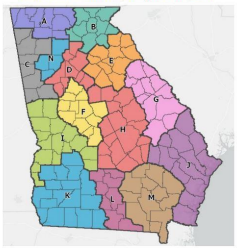


Adult Ventilator Statistics	
Capacity	2,807
In Use	855 (30%)
Availability	1,952 (70%)

Medical Facility Essential Elements

	ER Beds			Critical Care Beds			General Inpatient Beds		
	Cap	*Use*	*Avail*	*Cap*	*Use*	*Avail*	*Cap*	*Use*	*Avail*
Region A	133	33	100	46	30	16	464	215	249
Region B	185	34	151	147	88	59	762	557	205
Region C	224	107	117	174	98	76	838	739	99
Region D	1,007	545	462	1,203	914	289	4,649	3,664	985
Region E	142	34	108	70	62	8	633	519	114
Region F	304	133	171	230	204	26	1,504	1,217	287
Region G	231	93	138	190	141	49	1,291	1,019	272
Region H	89	21	68	40	32	8	282	142	140
Region I	142	87	55	105	87	18	812	553	259
Region J	326	108	218	211	175	36	1,025	793	232
Region K	161	46	115	151	79	72	794	568	226
Region L	108	25	83	76	54	22	516	358	158
Region M	84	34	50	60	41	19	325	238	87
Region N	251	99	152	209	173	36	1,272	1,068	204
TOTALS	3,387	1,399	1,988	2,912	2,178	734	15,167	11,650	3,517

Regional Coordinating Hospitals Regions



Q: What is the total capacity of ER Beds in regions H and K combined?
A: 464

Figure 14. Examples from our dataset real image-synthetic text, where the sample is ambiguous or incorrect.

Synthetic image/Synthetic Text	
VizWiz	**OkVQA**
Q: Are there any people in the image? **A**: Unanswerable **Prompt:** An overexposed image with excessive lighting, washing out most of the details	**Q**:Which type of cookie is next to the oatmeal raisin cookie? When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase. **A**: Chocolate chip **Prompt**: A collection of different types of cookies (chocolate chip, macarons, oatmeal raisin) with clear visual details.

Figure 15. Examples from our dataset synthetic image-synthetic text, where the sample is ambiguous or incorrect. For readability, the ScienceQA image was cropped to focus on the region of interest.