Learning from Reasoning Failures via Synthetic Data Generation

Gabriela Ben Melech Stan[1], Estelle Aflalo[1], Avinash Madasu[1], Vasudev Lal[2], Phillip Howard[3]

[1]Intel Labs
[2]Oracle
[3]Thoughtworks
{gabriela.ben.melech.stan, estelle.aflalo, avinash.madasu}@intel.com, vasudev.lal@oracle.com,
phillip.howard@thoughtworks.com

Abstract

Training models on synthetic data has emerged as an increasingly important strategy for improving the performance of generative AI. This approach is particularly helpful for large multimodal models (LMMs) due to the relative scarcity of high-quality paired image-text data compared to language-only data. While a variety of methods have been proposed for generating large multimodal datasets, they do not tailor the synthetic data to address specific deficiencies in the reasoning abilities of LMMs which will be trained with the generated dataset. In contrast, humans often learn in a more efficient manner by seeking out examples related to the types of reasoning where they have failed previously. Inspired by this observation, we propose a new approach for synthetic data generation which is grounded in the analysis of an existing LMM's reasoning failures. Our methodology leverages frontier models to automatically analyze errors produced by a weaker LMM and propose new examples which can be used to correct the reasoning failure via additional training, which are then further filtered to ensure high quality. We generate a large multimodal instruction tuning dataset containing over 553k examples using our approach and conduct extensive experiments demonstrating its utility for improving the performance of LMMs on multiple downstream tasks. Our results show that models trained on our synthetic data can even exceed the performance of LMMs trained on an equivalent amount of additional 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) have significantly advanced the state-of-art across a broad range of problem domains, which can be attributed in large part to scaling the size of models and datasets used for training. However, a limiting factor to further scaling of training datasets is the availability of new high-quality data. With recent studies suggesting that we are rapidly approaching saturation of existing public data sources (Villalobos et al. 2022; Xue et al. 2023), synthetic data generation has become an increasingly popular approach for augmenting training datasets. The generation of high-quality synthetic data has become more feasible recently as the capabilities of LLMs have increased, thereby enabling their use for the production of synthetic data at scale (Ding et al. 2024; Yu et al. 2023; Lee et al. 2024; Howard et al. 2022, 2023; Su et al. 2024).

Synthetic data has been especially valuable for training large multimodal models (LMMs), which combine an LLM with a vision encoder to enable text generation conditioned on multimodal inputs. Unlike LLMs, real data for training LMMs is scarce due to the lack of naturally-occurring images paired with high-quality text. Consequently, popular LMMs such as LLaVA (Liu et al. 2023) use partially synthetic data for multimodal instruction tuning, which is often collected by generating synthetic text for real images.

Existing approaches to generating synthetic data suitable for training LMMs suffer from two primary limitations. First, most rely on the acquisition of real images, which are then paired with synthetic text produced by another LMM. This limits their application to domains where images are readily available and therefore makes such approaches ill-suited for low-resource settings where image data is scarce. Second, prior approaches utilize broad generation strategies which arbitrarily produce synthetic examples without consideration of the type of data which would be most useful for improving LMMs. This results in an inefficient approach for both generating data and training LMMs, as many of the synthetic examples may provide little or no incremental value when added to training datasets.

In contrast to the dominant paradigm for generating synthetic training data, humans often learn more efficiently by focusing their attention on examples which are most related to their past reasoning failures. To acquire expert knowledge, Ericsson, Krampe, and Tesch-Römer (1993) argue that humans should "ideally be given explicit instructions about the best method and be supervised by a teacher to allow individualized diagnosis of errors, informative feedback, and remedial part training." Since humans acquire new knowledge by learning from their failures (Darabi, Arrington, and Sayilir 2018; Brown, Collins, and Duguid 1989; Morgan 1904), seeking out additional examples related to past reasoning errors for further practice and evaluation helps the learner achieve expertise in a subject. Problems which require types of reasoning that have already been mastered are often ignored under this paradigm, as they provide little incremental value in the learning process.

Motivated by such observations of human learning be-

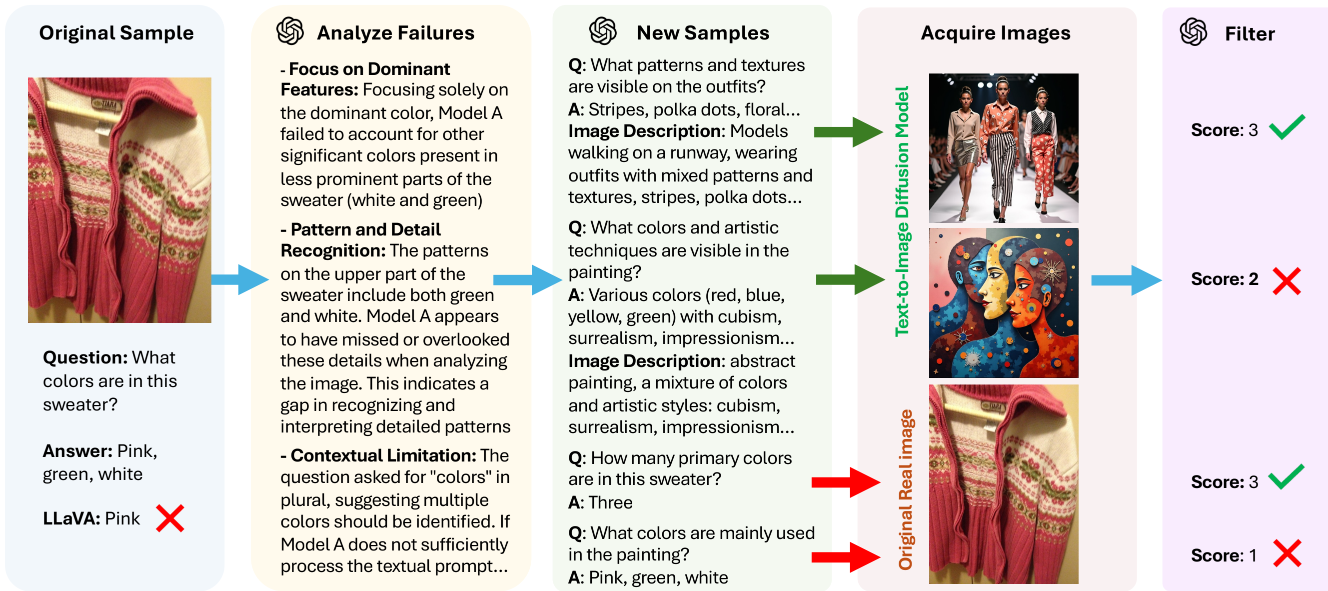

Figure 1: Illustration of our approach. Given a sample from an existing dataset which LLaVA answers incorrectly, we prompt a frontier model to analyze LLaVA's reasoning failures and propose new synthetic samples which require similar types of reasoning.

havior, we propose a new approach for generating synthetic multimodal data which is grounded in an existing LMM's reasoning failures. First, we identify the failure modes of an existing LMM by evaluating it on the training split of a benchmark dataset and then asking a strong frontier model to hypothesize possible reasons for its incorrect predictions. The frontier model then proposes new question & answer pairs related to the identified reasoning failure modes as well as descriptions of images to accompany them, which can be either used to link the new examples to existing images, or as prompts to generate synthetic images. We further filter the resulting synthetic data using an LMM-as-a-judge methodology to ensure high quality.

Using our approach, we generate a large multimodal instruction tuning dataset containing over 553k examples derived from the reasoning failures of LLaVA-1.5-7B, a popular LMM. We conduct extensive training experiments showing that our synthetic data improves LLaVA's performance on multiple downstream tasks, even outperforming additional training on alternative real datasets. This is notable, as prior studies have shown that training on synthetic data often requires significantly more examples to achieve equivalent results as training on real data (He et al. 2022). Training on our dataset also produces greater performance gains than previously-proposed synthetic datasets, highlighting the utility of grounding synthetic data generation in the analysis of a model's reasoning failures.
To summarize, our contributions are as follows:

1. We propose a new approach for generating fully synthetic multimodal datasets which is grounded in an analysis of an existing LMM's reasoning failures.
2. Using our methodology, we produce a large multimodal instruction tuning dataset containing over 553k examples derived from errors produced by LLaVA.
3. We conduct extensive experiments demonstrating the utility of our dataset for improving LMMs via training.
4. Through ablations and human analyses, we demonstrate the high quality of data produced by our approach.

2 Related Work

Synthetic Datasets for Training LMMs. Chen et al. (2024b) proposed the ALLaVA dataset, which consists of 1.3M samples of real images paired with image annotations and question-answer pairs generated by a frontier LMM. While their dataset is similarly suited for multimodal instruction tuning as our dataset, they utilize only real images and do not ground data generation in model failures, whereas our approach can utilize both real and synthetically generated images. Li et al. (2024) generate synthetic question-answer pairs from an LLM for real chart images to address the narrower domain of reasoning-based chart VQA. Yang et al. (2025) generate synthetic multimodal data for text-rich images, leveraging code-guided generation in languages such as LaTeX and HTML to automatically generate synthetic images with dense text descriptions. In contrast, our approach can be broadly applied to different problem domains while leveraging text-to-image diffusion models to produce more diverse types of images.

Other approaches to generating synthetic data suitable for training MLLMs differ from ours in the types of reasoning which they target. Img-Diff (Jiao et al. 2025) contains contrasting synthetic examples which differ only in terms of certain objects in specific regions of an image, requiring the MLLM to learn to identify and articulate such differences. Data-Juicer 2.0 (Chen et al. 2024a) provides multimodal synthetic data generation capabilities utilizing foundation MLLMs and text-to-image diffusion models, but does not address our specific task of generating synthetic data derived from reasoning failures. For a comprehensive data-centric review of MLLMs and how they can assist in the data generation process, see Qin et al. (2025).

Synthetic Data Generation from Model Failures. Some prior work has explored the use of model failures in the context of synthetic data generation. Jain et al. (2022) proposed an approach to identify directions in a vision model's latent space which correspond to model failures, which can then be integrated into diffusion models to generate synthetic images tailored to correcting the model's mistakes. Chegini and Feizi (2023) use ChatGPT and CLIP to identify text descriptions corresponding to a vision model's failure modes, which are used as prompts for diffusion models to produce synthetic images for data augmentation. In the realm of language-only models, DISCERN (Menon and Srivastava 2024) utilizes two LLMs to iteratively produce natural language descriptions of errors for synthetic data generation. Lee et al. (2024) use incorrect answers from a student LLM finetuned on specific tasks as input to a teacher LLM which generates new examples to use for training. These prior works differ from ours in that they (1) generate only single-modality data (e.g., text-only or image-only), and (2) focus on classification tasks such as image recognition or text classification. In contrast, our approach generates image-text datasets aimed at training models for open-ended text generation conditioned on multimodal inputs.

Generating Synthetic Data from Frontier Models to Teach New Skills. AgentInstruct (Mitra et al. 2024) 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. (2024) 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 (Gupta et al. 2023; Ubani, Polat, and Nielsen 2023) as well as knowledge distillation from a teacher model (Kim, Jang, and Yang 2024) 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.

3 Dataset Construction

3.1 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, which we denote as Model Failure Sets (MFS). Each MFS highlights the weaknesses of the baseline LMM, providing a focused challenging dataset.

3.2 Synthetic Data Generation

Using the resulting MFS, we guide the frontier model through a structured multi-turn process to diagnose the failures of the baseline LMM and generate new training samples to address these failure modes.

The frontier model first analyzes the image, question, ground truth, and incorrect prediction to diagnose reasoning errors. It is then instructed to propose new challenging samples consisting of a detailed image description, question, and deterministic answer designed to address the failures. We explore two sourcing strategies for the images: real images and synthetically generated ones. These two methods utilize the prompt described in Figure 5 of the Appendix. For Method 1, the final image-generation step is skipped.

Method 1: Question-Answer Generation for Existing Images. We leverage the original images from failed samples and prompt the frontier model to generate 10 new question-answer pairs per sample using the reasoning pipeline described above. This approach is particularly effective for benchmarks like InfoVQA and ScienceQA, where text-to-image models often fail to capture fine-grained text and generate accurate spatial details.

Method 2: Fully Synthetic Question, Answer, and Image Generation. In this approach, the frontier model generates both a question-answer pair and a detailed image description, which is used to prompt a text-to-image diffusion model to generate synthetic images. For each failed sample, we create 100 fully synthetic samples, consisting of 10 (image prompt, question, answer) triplets, with each image prompt used to produce 10 images using different classifier-free guidance scales to enhance image diversity.

Encouraging Diversity. To further enhance the diversity of our data, we use a variation of our prompt which instructs the frontier LMM to "provide examples that challenge Model A's weaknesses using scenarios from entirely different domains." This additional instruction is appended to Step 4 (Figure 5) to encourage the generation of samples in different domains and enhance generalization. Figure 2 compares fully synthetic data with similar and non-similar domains. We also varied constraints on the question format (e.g., multiple-choice, true/false) and answer instructions (e.g., requiring "Unanswerable" or brief responses, see the Shiba Inu example in Figure 10).

Filtering We use the same frontier LMM to filter synthetic samples by rating image-question-answer triplets from 1 (incorrect) to 3 (fully correct), as shown in the prompt in Figure 6. Only samples rated 3 are retained to ensure quality.

3.3 Details of Generated Dataset

Our synthetic dataset consists of 553,992 samples incorporating both real and generated images derived from the MFS

Figure 2: 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.

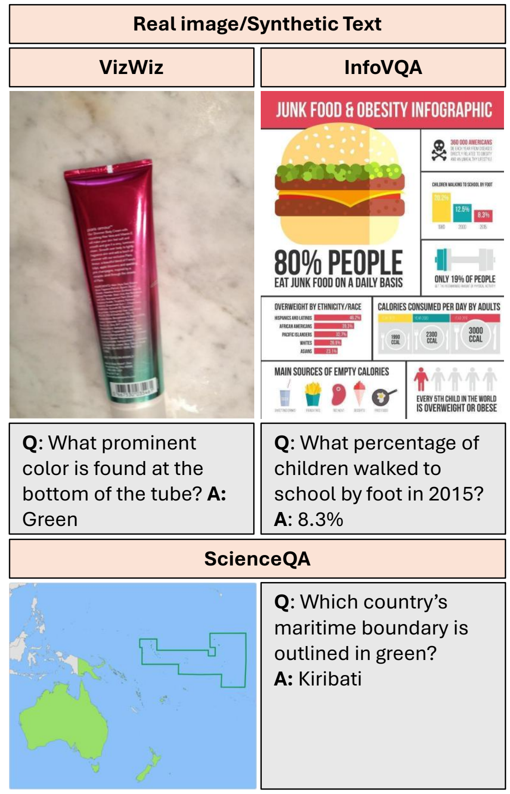

Figure 3: Examples of generated synthetic question-answer pairs for real images from VizWiz, InfoVQA, and ScienceQA.

of LLaVA-1.5-7B on four benchmark training sets: VizWiz (Gurari et al. 2018), InfoVQA (Mathew et al. 2022), ScienceQA (Lu et al. 2022), and OK-VQA (Marino et al. 2019). These benchmarks were chosen to cover a wide range of visual reasoning challenges. VizWiz features real-world images from visually impaired users, demanding detailed scene understanding. OK-VQA requires external knowledge for visual question answering. InfoVQA focuses on text-rich images, testing reading comprehension. ScienceQA involves multimodal scientific questions, assessing spatial and logical reasoning skills. We use GPT-4o (OpenAI 2024) as the frontier LMM for analyzing the reasoning failures of LLaVA-1.5-7B due to its strong multimodal reasoning capabilities. The text-to-image model FLUX.1-schnell (Labs 2024) was used to generate the images with a resolution of 1024×1024 pixels and a classifier guidance scale ranging from 3 to 13.

Table 6 (Appendix) provides an overview of each benchmark, including the size of the training split, the MFS size, and the number of synthetic samples after filtering. For InfoVQA and ScienceQA, we applied Method 1 from Section 3.2. For OK-VQA, we employed Method 2. For the VizWiz, we integrated both methods. The final dataset comprises 42% real-image-based and 58% fully synthetic samples, each with a single-turn question-answer pair.

Our filtering process effectively removes low-quality samples. VizWiz had the highest rejection rate (81% for synthetic images and 34% for real images), highlighting the challenge of generating quality synthetic visuals. OK-VQA's lower rate (29%) suggests simpler and less ambiguous visual content. ScienceQA showed a similar 29% rejection rate for real images, while InfoVQA had only 5% removed. Figure 3 showcases diverse question types and the quality of generated samples from VizWiz and InfoVQA. Appendix D includes additional examples, including samples with errors or ambiguities. Finally, Figure 8 shows OK-VQA examples with original image-question-answer pairs, the GPT-4o-identified failure mode, and newly generated samples targeting the same failure.

4 Experiments

4.1 Details of Experimental Setting

We perform vision-language instruction tuning with LLaVA-v1.5-7B, building upon the LLaVA-1.5-mix-665K instruction tuning dataset (Liu et al. 2024). This dataset contains 665K structured user-GPT conversations mostly focused on visual prompts. Because our approach explicitly targets image-grounded reasoning, we only selected the 624K samples that include images from the original LLaVA-instruct data. We then augmented this visually oriented subset with our generated synthetic MFS (Section 3.3), varying the number of synthetic samples used according to the requirements of each experiment. The resulting training set combines real image-text conversations and generated synthetic samples that address specific reasoning failures. For fine-tuning, we used Vicuna-v1.5-7B (Zheng et al. 2023) 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 (details in Appendix A.5). 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.

4.2 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, in OK-VQA, the original training set consists of 9,009 samples, but we utilized only 607 training samples where LLaVA failed 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. We also provide results for LLaVA models

Dataset	Model	N	N_{syn}	EM Score
I-VQA	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**
SQA	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 trained using synthetically augmented data ($\text{LLaVA}_{\text{syn}}$), and augmented with real in-domain data ($\text{LLaVA}_{\text{real}}$). I-VQA denotes InfoVQA and SQA denotes ScienceQA.

trained on four alternative synthetically generated datasets: ALLaVA (Chen et al. 2024b), CoSyn-400k (Yang et al. 2025), SimVQA (Cascante-Bonilla et al. 2022), and Img-Diff (Jiao et al. 2024)[1]. From Table 2, we observe that our synthetic data outperforms these baselines datasets, demonstrating the value of grounding synthetic data generation in an analysis of a model's failures. Finally, we evaluated trained models on OOD benchmarks (Table 8 of Appendix) and found that models trained on our dataset do not exhibit degradation in OOD performance, demonstrating the ability of our approach to correct model reasoning failures without sacrificing general reasoning capabilities. Importantly, our dataset strikes a middle ground in the cost–quality–diversity tradeoff and can be further complemented with lower-cost image generation pipelines such as CoSyn's to achieve scalable and semantically rich supervision. A detailed comparison of compute and generation costs across our method and baseline datasets is provided in Appendix B.3.

4.3 Generalization to Other Models and Tasks

While our synthetic data is collected by targeting the reasoning failures of a specific LMM, we also evaluated whether it could generalize to other models and tasks, thereby further increasing its value beyond correcting targeted reasoning failures in a single model. To evaluate this generalization ability of our synthetic dataset, we trained models using the same datasets as before but with different backbone LLMs. In the following experiments, we used Gemma-2B (Team et al. 2024) and Qwen2-7B (Yang et al. 2024) as base LLMs. We adopted the LLaVA two-phase training procedure: pretraining on the LLaVA 558k dataset (Liu et al. 2024) followed by instruction fine-tuning. Table 2 indicates

[1]ALLaVA consists of real images with synthetic text, whereas Cosyn-400k and SimVQA contain fully synthetic examples. Our synthetic data utilized here contains both real and synthetic images.

Base LLM	N	N_{syn}	Aug Data	I-VQA	OK-VQA	SQA
Vicuna-7B	625k	0	N/A (bs)	26.7	57.0	70.7
Vicuna-7B	643k	18k	Img-Diff	27.3	56	71.3
			Ours	**31.5**	**61.2**	**72**
Vicuna-7B	687k	62k	SimVQA	26.6	54.4	71.1
			Ours	**33.1**	**60.8**	**73.1**
Vicuna-7B	750k	125k	ALLaVA	28.8	49.4	66.5
			CoSyn	29.6	57.7	70.9
			Ours	**33.2**	**61.1**	**73.0**
Gemma-2B	625k	0	Ours	21.8	51.7	62.3
	749k	125k		**29.8**	**54.8**	**65.2**
Qwen2-7B	625k	0	Ours	26.7	59.2	77.9
	750k	125k		**27**	**60.6**	**80.7**

Table 2: **Training data augmentation 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. I-VQA denotes InfoVQA and SQA represents ScienceQA. 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.

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 across nearly all benchmarks. Notably, despite being generated based on LLaVA-Vicuna-7B failure modes, our synthetic data enhances the performance of other models whether they are of the same size (Qwen2-7B) or smaller (Gemma-2B).

For some models (e.g., Qwen2.5-VL-7B-Instruct), the original finetuning dataset is not publicly available. Nevertheless, we show that our dataset can be sufficient to improve these models through continued finetuning. Table 4 provides the results of finetuning Qwen2.5-VL-7B on our 124K synthetic-only dataset derived from LLaVA reasoning failures. We observe that finetuning Qwen2.5-VL-7B on our 124K dataset improves performance significantly on OK-VQA and ScienceQA. InfoVQA performance remains unchanged, likely due to Qwen's high baseline and prior exposure to similar data. Generating synthetic data which targets Qwen-specific failure cases may yield further gains; nevertheless, these results point to the value of our dataset beyond correcting reasoning failures in only the targeted LMM.

Finally, we investigated whether our targeted synthetic dataset could improve performance on other tasks besides those used for deriving the reasoning failures. Specifically, we evaluated our $\text{LLaVA}_{\text{syn}}$ model trained on the 124K synthetic dataset on multiple hallucination and math benchmarks. The results in Table 3 show that $\text{LLaVA}_{\text{syn}}$ outperforms the baseline LLavA model across all such benchmarks, demonstrating how our synthetic dataset improves model performance beyond the specific reasoning failures targeted during the dataset generation process. The ability of

	LLaVA	$\text{LLaVA}_{\text{syn}}$ (124k)
HalluB↑	40.6	**47.4**
LLaVA-B-COCO↑	85.3	**86.9**
POPE↑	86.8	**87.2**
Math-V(vision)↑	10.6	**11.7**
MathVista↑	24.9	**26.6**

Table 3: Our targeted 124k synthetic dataset generalizes to improve LLaVA on hallucination and math benchmarks.

	InfoVQA	OK-VQA	SciQA
LLaVA1.5-7b	26.4	**57.1**	70.8
+ CFT	**32.4**	45.8	**73.2**
Qwen2.5-VL-7b-Instruct	82.6	42.4	67.9
+ CFT	82.5	**53.0**	**85.5**

Table 4: Continued finetuning results for LLaVA-1.5-7b and Qwen2.5-VL-7B-Instruct with our 124K synthetic dataset

our synthetic dataset to generalize to improving other LMMs and tasks than the ones used in the generation process further illustrates the significant value of our approach.

4.4 Continued Finetuning Results

We also conducted experiments wherein we continued finetuning LLaVA on our 124K mixture dataset. This approach allows us to evaluate performance gains attributed solely to our data while significantly reducing compute requirements compared to full finetuning. In this setting, we used the same hyperparameter configurations as our previous finetuning experiments. Table 4 shows that continued finetuning improves in-domain performance except on OK-VQA. Notably, continued finetuning on the real OK-VQA training set (9K, 41.1%) results in worse performance than with an identical-size synthetic dataset (48.2%), suggesting possible overfitting when training on real data.

4.5 Correcting Specific Types of Reasoning Failures

Our synthetic data generation approach identifies specific types of LMM failures. To systematically categorize these

Cluster name	Percentage (%)
Feature misinterpretation	8
Blurriness	12.4
Weakness in visual analysis	18.6
Incomplete context understanding	24
Other	11.3
Text recognition errors	9.2
Object recognition failure	10.8
Overgeneralization	5.6

Table 5: Table shows the clusters of LLAVA reasoning failures described by GPT-4o.

Dataset	LLAVA	$\text{LLaVA}_{\text{syn}}$
CIFAR-10	**82.1**	81.2
Food-101	**13.4**	13.2
iNaturalist	20.6	**52.0**
MNIST	75.1	**80.5**
F-MNIST	9.8	**10.0**
Oxford-pets	39.6	**96.4**

Table 6: Image classification results, with $\text{LLaVA}_{\text{syn}}$ augmented only using examples for object recognition failures.

failures, we encoded each reasoning explanation using sentence transformers (Reimers and Gurevych 2019) and clustered them using k-means. Table 5 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 stage of LLaVA finetuning. The model was then evaluated on CIFAR-10 (Krizhevsky, Hinton et al. 2009), Food-101 (Bossard, Guillaumin, and Van Gool 2014), iNaturalist (Van Horn et al. 2018), MNIST (LeCun et al. 1998), Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), and Oxford-Pets (Binary) (Parkhi et al. 2012) by formatting samples as multiple-choice questions. Table 6 presents a comparison of LLaVA and $\text{LLaVA}_{\text{syn}}$. $\text{LLaVA}_{\text{syn}}$ surpasses LLaVA on four out of six datasets, with significant improvements on iNaturalist, MNIST and Oxford-Pets. This demonstrates the effectiveness of our synthetic dataset in addressing specific reasoning failures within LLaVA. 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.

5 Analysis

5.1 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 semantic coherence, 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. Additionally, to assess the frontier model's reasoning and ensure it highlights meaningful failures, five annotators rated 209 samples (image, QA, and error analysis) on a 1–3 scale. The average score of 2.81 indicates that the explanations usually identified the true failure cause.

5.2 Impact of LMM on Generated Data

We evaluate our synthetic-data pipeline with two LMMs of very different cost profiles: the proprietary GPT-4o API and the fully open, Apache-2.0-licensed Qwen2-VL-7B (Wang et al. 2024). Both models improve LLaVA-7B after fine-tuning on the resulting data: GPT-4o yields gains of 2.7% on InfoVQA and 10% on OK-VQA, while Qwen2-VL-7B delivers 7.0% and 3.5% on the same tasks. Although GPT-4o's detailed and precise reasoning may contribute to generating more targeted and effective synthetic samples, which produces larger gains on some benchmarks, our failure-guided generation shows robustness by producing meaningful improvements across both models. Because Qwen2-VL-7B runs locally on a single A100 GPU, synthetic data can be obtained with our approach at a lower cost than commercial models such as GPT-4o. Figure 7 (Section C.1 in Appendix) illustrates representative examples of the reasoning and sample diversity across the two models.

6 Conclusion

We introduced a novel method for generating multimodal synthetic data by analyzing model reasoning failures. This approach enabled us to create a multimodal instruction tuning dataset with over 553k synthetic examples derived from LLaVA's errors. Experimental results show that our synthetic data significantly improves LLaVA's in-domain performance on InfoVQA, ScienceQA, and OK-VQA, even outperforming training on an equivalent amount of real data sampled from most of these datasets. Furthermore, models trained on our synthetic dataset exhibit improvements in OOD evaluations and outperform training on other existing synthetic datasets when the amount of training data augmentation is scaled. 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. Through additional ablations and human evaluations, we demonstrated the effectiveness of different components of our methodology and the high quality of our dataset. We believe our study points to the promise of targeted synthetic data generation strategies which leverage an understanding of a model's reasoning deficiencies to construct better synthetic examples.

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. 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, 446–461. Springer.

Brown, J. S.; Collins, A.; and Duguid, P. 1989. Situated cognition and the culture of learning. *1989*, 18(1): 32–42.

Cascante-Bonilla, P.; Wu, H.; Wang, L.; Feris, R.; and Ordonez, V. 2022. SimVQA: Exploring Simulated Environments for Visual Question Answering. arXiv:2203.17219.

Chegini, A.; and Feizi, S. 2023. Identifying and Mitigating Model Failures through Few-shot CLIP-aided Diffusion Generation. *arXiv preprint arXiv:2312.05464*.

Chen, D.; Huang, Y.; Pan, X.; Jiang, N.; Wang, H.; Zhang, Y.; Ge, C.; Chen, Y.; Zhang, W.; Ma, Z.; et al. 2024a. Data-Juicer 2.0: Cloud-Scale Adaptive Data Processing for and with Foundation Models. *arXiv preprint arXiv:2501.14755*.

Chen, G. H.; Chen, S.; Zhang, R.; Chen, J.; Wu, X.; Zhang, Z.; Chen, Z.; Li, J.; Wan, X.; and Wang, B. 2024b. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*.

Darabi, A.; Arrington, T. L.; and Sayilir, E. 2018. Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development*, 66: 1101–1118.

Ding, B.; Qin, C.; Zhao, R.; Luo, T.; Li, X.; Chen, G.; Xia, W.; Hu, J.; Tuan, L. A.; and Joty, S. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, 1679–1705.

Ericsson, K. A.; Krampe, R. T.; and Tesch-Römer, C. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3): 363.

Gupta, H.; Scaria, K.; Anantheswaran, U.; Verma, S.; Parmar, M.; Sawant, S. A.; Baral, C.; and Mishra, S. 2023. Targen: Targeted data generation with large language models. *arXiv preprint arXiv:2310.17876*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.

He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P.; Bai, S.; and Qi, X. 2022. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.

Howard, P.; Singer, G.; Lal, V.; Choi, Y.; and Swayamdipta, S. 2022. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. *arXiv preprint arXiv:2210.12365*.

Howard, P.; Wang, J.; Lal, V.; Singer, G.; Choi, Y.; and Swayamdipta, S. 2023. Neurocomparatives: Neuro-symbolic distillation of comparative knowledge. *arXiv preprint arXiv:2305.04978*.

Jain, S.; Lawrence, H.; Moitra, A.; and Madry, A. 2022. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*.

Jiao, Q.; Chen, D.; Huang, Y.; Ding, B.; Li, Y.; and Shen, Y. 2024. Img-Diff: Contrastive Data Synthesis for Multimodal Large Language Models. arXiv:2408.04594.

Jiao, Q.; Chen, D.; Huang, Y.; Ding, B.; Li, Y.; and Shen, Y. 2025. Img-diff: Contrastive data synthesis for multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9296–9307.

Kim, G.; Jang, D.; and Yang, E. 2024. Promptkd: Distilling student-friendly knowledge for generative language models via prompt tuning. *arXiv preprint arXiv:2402.12842*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lee, N.; Wattanawong, T.; Kim, S.; Mangalam, K.; Shen, S.; Anumanchipalli, G.; Mahoney, M. W.; Keutzer, K.; and Gholami, A. 2024. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*.

Li, Z.; Jasani, B.; Tang, P.; and Ghadar, S. 2024. 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*, 13613–13623.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

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

Mathew, M.; Bagal, V.; Tito, R.; Karatzas, D.; Valveny, E.; and Jawahar, C. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1697–1706.

Menon, R. R.; and Srivastava, S. 2024. DISCERN: Decoding Systematic Errors in Natural Language for Text Classifiers. *arXiv preprint arXiv:2410.22239*.

Mitra, A.; Del Corro, L.; Zheng, G.; Mahajan, S.; Rouhana, D.; Codas, A.; Lu, Y.; Chen, W.-g.; Vrousgos, O.; Rosset, C.; et al. 2024. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*.

Morgan, C. L. 1904. *An introduction to comparative psychology*. Walter Scott Publishing Company.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Qin, Z.; Chen, D.; Zhang, W.; Yao, L.; Huang, Y.; Ding, B.; Li, Y.; and Deng, S. 2025. The synergy between data and multi-modal large language models: A survey from co-development perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Reimers, N.; and Gurevych, I. 2019. 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.

Su, X.; Luo, M.; Pan, K. W.; Chou, T. P.; Lal, V.; and Howard, P. 2024. SK-VQA: Synthetic Knowledge Generation at Scale for Training Context-Augmented Multimodal LLMs. *arXiv preprint arXiv:2406.19593*.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; Sessa, P. G.; Chowdhery, A.; Roberts, A.; Barua, A.; Botev, A.; Castro-Ros, A.; Slone, A.; Héliou, A.; Tacchetti, A.; Bulanova, A.; Paterson, A.; Tsai, B.; Shahriari, B.; Lan, C. L.; Choquette-Choo, C. A.; Crepy, C.; Cer, D.; Ippolito, D.; Reid, D.; Buchatskaya, E.; Ni, E.; Noland, E.; Yan, G.; Tucker, G.; Muraru, G.-C.; Rozhdestvenskiy, G.; Michalewski, H.; Tenney, I.; Grishchenko, I.; Austin, J.; Keeling, J.; Labanowski, J.; Lespiau, J.-B.; Stanway, J.; Brennan, J.; Chen, J.; Ferret, J.; Chiu, J.; Mao-Jones, J.; Lee, K.; Yu, K.; Millican, K.; Sjoesund, L. L.; Lee, L.; Dixon, L.; Reid, M.; Mikuła, M.; Wirth, M.; Sharman, M.; Chinaev, N.; Thain, N.; Bachem, O.; Chang, O.; Wahltinez, O.; Bailey, P.; Michel, P.; Yotov, P.; Chaabouni, R.; Comanescu, R.; Jana, R.; Anil, R.; McIlroy, R.; Liu, R.; Mullins, R.; Smith, S. L.; Borgeaud, S.; Girgin, S.; Douglas, S.; Pandya, S.; Shakeri, S.; De, S.; Klimenko, T.; Hennigan, T.; Feinberg, V.; Stokowiec, W.; hui Chen, Y.; Ahmed, Z.; Gong, Z.; Warkentin, T.; Peran, L.; Giang, M.; Farabet, C.; Vinyals, O.; Dean, J.; Kavukcuoglu, K.; Hassabis, D.; Ghahramani, Z.; Eck, D.; Barral, J.; Pereira, F.; Collins, E.; Joulin, A.; Fiedel, N.; Senter, E.; Andreev, A.; and Kenealy, K. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295.

Ubani, S.; Polat, S. O.; and Nielsen, R. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Villalobos, P.; Sevilla, J.; Heim, L.; Besiroglu, T.; Hobbhahn, M.; and Ho, A. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 1.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xue, F.; Fu, Y.; Zhou, W.; Zheng, Z.; and You, Y. 2023. To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36: 59304–59322.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Yang, Y.; Patel, A.; Deitke, M.; Gupta, T.; Weihs, L.; Head, A.; Yatskar, M.; Callison-Burch, C.; Krishna, R.; Kembhavi, A.; et al. 2025. Scaling Text-Rich Image Understanding via Code-Guided Synthetic Multimodal Data Generation. *arXiv preprint arXiv:2502.14846*.

Yu, Y.; Zhuang, Y.; Zhang, J.; Meng, Y.; Ratner, A. J.; Krishna, R.; Shen, J.; and Zhang, C. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36: 55734–55784.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Ziegler, I.; Köksal, A.; Elliott, D.; and Schütze, H. 2024. CRAFT Your Dataset: Task-Specific Synthetic Dataset Generation Through Corpus Retrieval and Augmentation. *arXiv preprint arXiv:2409.02098*.