HONEYBEE: Data Recipes for Vision-Language Reasoners

Hritik Bansal[1,2*] Devendra Singh Sachan[1] Kai-Wei Chang[2] Aditya Grover[2]
Gargi Ghosh[1] Wen-tau Yih[1] Ramakanth Pasunuru[1]
[1]FAIR at Meta [2]University of California Los Angeles
rpasunuru@meta.com

Abstract

Recent advances in vision-language models (VLMs) have made them highly effective at reasoning tasks. However, the principles underlying the construction of performant VL reasoning training datasets remain poorly understood. In this work, we introduce several data curation approaches and study their impacts on VL reasoning capabilities by carefully controlling training and evaluation setups. We analyze the effects of context (image and question pair) sources, implement targeted data interventions, and explore scaling up images, questions, and chain-of-thought (CoT) solutions. Our findings reveal that (a) context source strategies significantly affect VLM performance, (b) interventions such as auxiliary signals from image captions and the inclusion of text-only reasoning yield substantial gains, and (c) scaling all data dimensions (e.g., unique questions per image and unique CoTs per image-question pair) consistently improves reasoning capability. Motivated by these insights, we introduce HONEYBEE*, a large-scale, high-quality CoT reasoning dataset with* 2.5*M examples consisting* 350*K image-question pairs. VLMs trained with* HONEYBEE *outperform state-of-the-art models across model sizes. For instance, a* HONEYBEE*-trained VLM with* 3*B parameters outperforms the SOTA model and the base model by* 7.8% *and* 24.8%*, respectively, on MathVerse. Furthermore, we propose a test-time scaling strategy that reduces decoding cost by* 73% *without sacrificing accuracy. Overall, this work presents improved strategies for VL reasoning dataset curation research.*[1]

1. Introduction

Solving reasoning problems, such as those involving mathematics in visual contexts, is a crucial capability for AI models, powering many real-world applications such as visual data analysis [39, 42], education [5], and scientific discovery [61]. Recent vision-language models (VLMs), such as GPT-4o [22], o3 [47], Gemini-2.5 [13], Llama-4 [44] achieve strong VL reasoning performance by training their pretrained models on high-quality synthetic chain-of-thought (CoT) data [75]. In particular, the ability to generate CoTs (e.g., step-by-step solution) during problem-solving enable VLMs to utilize additional inference-time computation before providing the final answer [19, 27, 69]. Yet, the multimodal CoT datasets and their recipes used for training state-of-the-art VLMs are often proprietary, leaving several open questions about their design space.

Several prior works have shown that supervised finetuning (SFT) with the quality of CoT data is crucial for LLM (text-only) reasoning performance [6, 18, 46, 58], and also serves as a key foundation for subsequent RL training [35]. However, there is a major gap in our understanding of how high-quality CoT datasets are constructed for VL reasoning, where a model needs to integrate information from several modalities (visual content in images and text content in questions) to provide accurate answers. Specifically, prior work on VL reasoning does not explore the breadth of design choices and suffers from several challenges. Firstly, the impact of context (image and question pairs) from diverse data sources remains unclear. For instance, `Math-LLaVA` [53] and `LLaVA-CoT` [71] curate different data distributions from existing image QA datasets and employ their own custom CoT generation strategies. Thus, it is uncertain how much of the reasoning performance of models trained on these datasets can be attributed to the quality of the context. Secondly, prior work [9, 20, 32] suggests that targeted data interventions—such as visual perturbations and difficulty filtering—can enhance model perception and problem-solving. However, there has been limited exploration of other interventions that could further improve data quality. Thirdly, while scaling up the training data is known to enhance reasoning performance [18], it remains unclear whether VL reasoning data should be scaled along specific axes such as the number of questions per image, and the number of CoTs per (image, question) pair.

Importantly, there is a lack of fair and robust compar-

*Work done at Meta.
[1]Data is available at https://huggingface.co/datasets/facebook/HoneyBee.

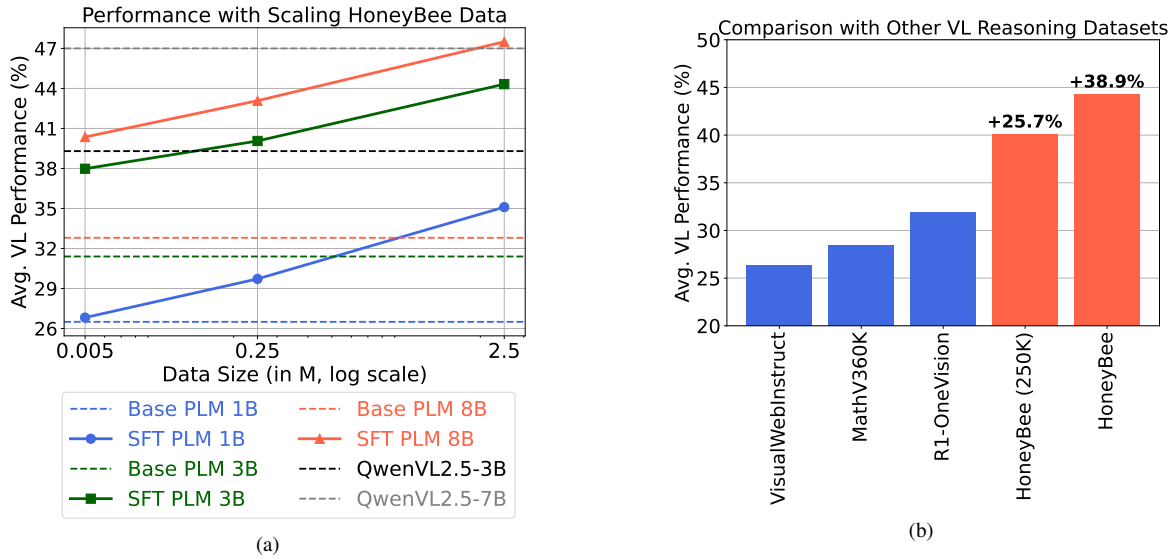


Figure 1. **Summary of the results.** (a) We show that training with increasing amounts of our curated HONEYBEE data leads to consistent accuracy improvements averaged across five VL reasoning tasks (MathVista, MathVerse, MathVision, MMMU-Pro, and We-Math) for several model sizes (1B to 8B). (b) We report the relative gains achieved by training HONEYBEE with PLM-3B compared to existing VL reasoning CoT datasets containing 250K–300K VL instances.

isons between diverse design decisions. For fairness, the direct effect of a design choice should be measured by fixing the training setups (e.g., SFT starting from identical models) and the evaluation protocol. For robustness, models should be trained at multiple scales (such as 3B and 8B) and evaluated on a collection of datasets rather than a single dataset. To address these critical questions in VL reasoning data design, we adopt a comprehensive and scalable approach to identify the key factors in dataset curation (Figure 5). This effort culminates in the creation of **HONEYBEE**, a high-quality and one of the largest VL reasoning datasets, comprising **2.5M** examples.

In our data curation process, we first study the impact of diverse contexts (i.e., image and question pairs) acquired from several VL reasoning datasets and rank them based on the performance of multiple VLMs (i.e., 3B and 8B models) on a battery of evaluation datasets. Interestingly, we observe that the choice of source datasets can lead to significant differences in model performance, with up to a 4% difference in average accuracy across evaluation datasets (§3.1). In the next stage, we create a list of data enhancement strategies, including *visual perturbation*, *text-rich images*, *perceptual redundancy*, *shallow perception*, *caption-and-solve*, *text-only reasoning*, *increased distractors*, *length* and *difficulty* filtering, applied on top of the best-performing dataset from the previous stage (Figure 2). Our experiments show that several data interventions fail to outperform the baseline dataset, highlighting their limited practical value despite strong motivations. Importantly, we reveal that auxiliary signals from image captions (*caption-and-solve*) and augmenting the VL reasoning data with text-only reasoning data lead to major improvements across diverse benchmarks (§5.2). In our experiments, we find that model performance improves with scaling all dimensions (images, questions, and CoTs) in the reasoning data (§5.3).

Motivated by these findings, we construct the HONEYBEE dataset, and train VLMs of several sizes (1B–8B) on the HONEYBEE dataset. We show that reasoning performance strongly scales with the amount of training data and outperforms existing state-of-the-art instruction-tuned VLMs, indicating at the quality and scalability of our dataset (§5.4). In particular, PLM-HONEYBEE-1B achieves a relative performance improvement of 28 percentage points (pp) over InternVL-3-1B-Instruct, averaged across ten evaluation datasets (Table 1). Furthermore, PLM-HONEYBEE-3B and PLM-HONEYBEE-8B achieve relative gains of 8.4pp and 2.7pp over Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-8B-Instruct, respectively. Ultimately, we also propose an efficient decoding strategy for test-time scaling that enables the generation of multiple solutions from the HONEYBEE-trained VL reasoners, using 73% fewer inference tokens without any loss in performance (§ 5.5). Our experiments yield insightful findings that lay the foundation for curating the next generation of VL reasoning datasets.

2. Preliminaries

In this work, we focus on the curation of high-quality, large-scale synthetic data to enable strong vision-language rea-

soning capabilities. Let $\mathcal{D} = \{(I_j, Q_j, A_j)\}_{j=1}^{N}$ denote the source vision-language reasoning dataset of size N, where each entry consists of an image I_j, a corresponding question Q_j, and an optional final answer A_j.[2] Further, let $\mathcal{G}$ be a synthetic data generator that outputs a textual chain-of-thought (CoT) to solve the questions about the images in $\mathcal{D}$, such that $C_j = \mathcal{G}(I_j, Q_j)$. In particular, the CoT C_j is composed of several reasoning steps, including step-by-step solutions, planning, self-verification, and self-reflection behaviors [57], denoted as S_j. This is followed by a predicted final answer P_j to the given question about the image. Thus, $C_j = [S_j; P_j]$, where ; denotes concatenation in the raw text space. Thus, We represent the synthetic data as $\mathcal{D}_{\mathcal{G}} = \{(I_j, Q_j, C_j, A_j)\}_{j=1}^{N}$.

Given synthetic data, we train a VLM (p_θ), using a supervised finetuning objective: $\mathbb{E}_{(I_j,Q_j,C_j)\in\mathcal{D}_{\mathcal{G}}}[\log p_\theta(C_j \mid I_j, Q_j)]$, i.e., maximizing the probability of the problem-solving CoT given the image and question as context. Post-training, the VLM will perform step-by-step reasoning before generating its predicted answer, thus utilizing additional test-time compute [69]. Overall, the goal of high-quality synthetic data is to train performant VL reasoners that can solve novel problems from diverse downstream tasks such as geometry, function plots, and charts [39]. In this work, we focus on the paradigm of training on smaller VLMs on synthetic data from larger generator models. This approach is more compute-efficient, more popular [6, 18, 30, 53], and allows for comprehensive training runs on diverse synthetic data distributions. In practice, p_θ can be a weaker VLM reasoner than the generator [53], a setup commonly referred to as knowledge distillation [21], where a strong teacher guides a weak student. Alternatively, the generator itself can be used as the student, a process known as self-improvement [54]. Prior research has also explored training a stronger reasoner with data from a weaker one, referred to as weak-to-strong reasoning [4, 8, 72].

3. Data Curation Pipeline

We outline our vision-language reasoning data curation pipeline (Appendix Figure 5), which consists of multiple stages: (a) `context curation`, which assesses the impact of diverse context (image, question) data sources, as well as their mixing (§3.1); (b) `data interventions`, which aim to enhance perception and problem-solving skills to enable strong VL reasoning capabilities (§3.2); and (c) `scaling`, which studies the impact of scaling diverse components of the reasoning data (§3.3).

3.1. Context Curation

The quality of context i.e., image, question pair, is crucial for determining the reasoning skills imparted to the VL reasoner [18]. For instance, a VL reasoner exposed to diverse geometric images and questions will excel in downstream geometry problem-solving [77]. In this work, we consider two stages of context curation: (a) **sourcing**, where we analyze the direct effect of the contexts (image, question pair) in individual datasets on training performant VL reasoners, and (b) **mixing**, where we combine the strengths of the top-performing data sources. Specifically, we mix the data from the top-2, top-4, and all data from the previous stage. We provide more details in Appendix 8.

3.2. Data Interventions

Starting with the best data mixture, we assess whether its quality can be further enhanced through targeted data interventions (Figure2). Specifically, these interventions aim to improve particular skills of the VL reasoner, including *perception* and *problem-solving* capabilities. Enhanced perception is crucial for robust visual understanding of image content within the model's context. Furthermore, strong problem-solving ability is essential for accurate calculations, planning, and reflection when addressing complex questions. The data intervention strategies can affect the original VL CoT reasoning data in various ways, including *replacement*[41], *augmentation*[3], and *filtering*[16].[3]

Perception Enhancement. We briefly outline various data interventions for enhancing the *perception* of the VL reasoner: (a) **visual perturbation**, where we create a perturbed version of the original image (*rotation*, *distractor concatenation*, *dominance-preserving mixup*) in the reasoning dataset to enhance the perceptual robustness of the VLM[32]; (b) **text-rich images**, where we aim to improve the model's ability to integrate textual information within images[63] by programmatically embedding the question and original image on a blank background of varying colors to generate new images; (c) **perceptual redundancy**, where we *filter* the original synthetic dataset by removing instances where the generator model leads to the correct final answer without access to the image, encouraging greater reliance on visual inputs; (d) **shallow perception**, where we filter the original dataset to remove instances where the generator model leads to the correct final answer with access to the question and image caption but not the image; and (e) **caption and solve**, where we enhance the visual understanding capabilities of the VL reasoner by providing auxiliary visual signals from the image caption generated by the model. We provide more details in Appendix 9.

[2] When there is no final answer, we can set $A_j = \phi$.

[3] The replacement strategy swaps some or all of the original data with higher-quality versions, keeping the dataset size unchanged. Augmentation increases data by adding transformed or external examples. Filtering removes low-quality instances based on set rules or classifiers.

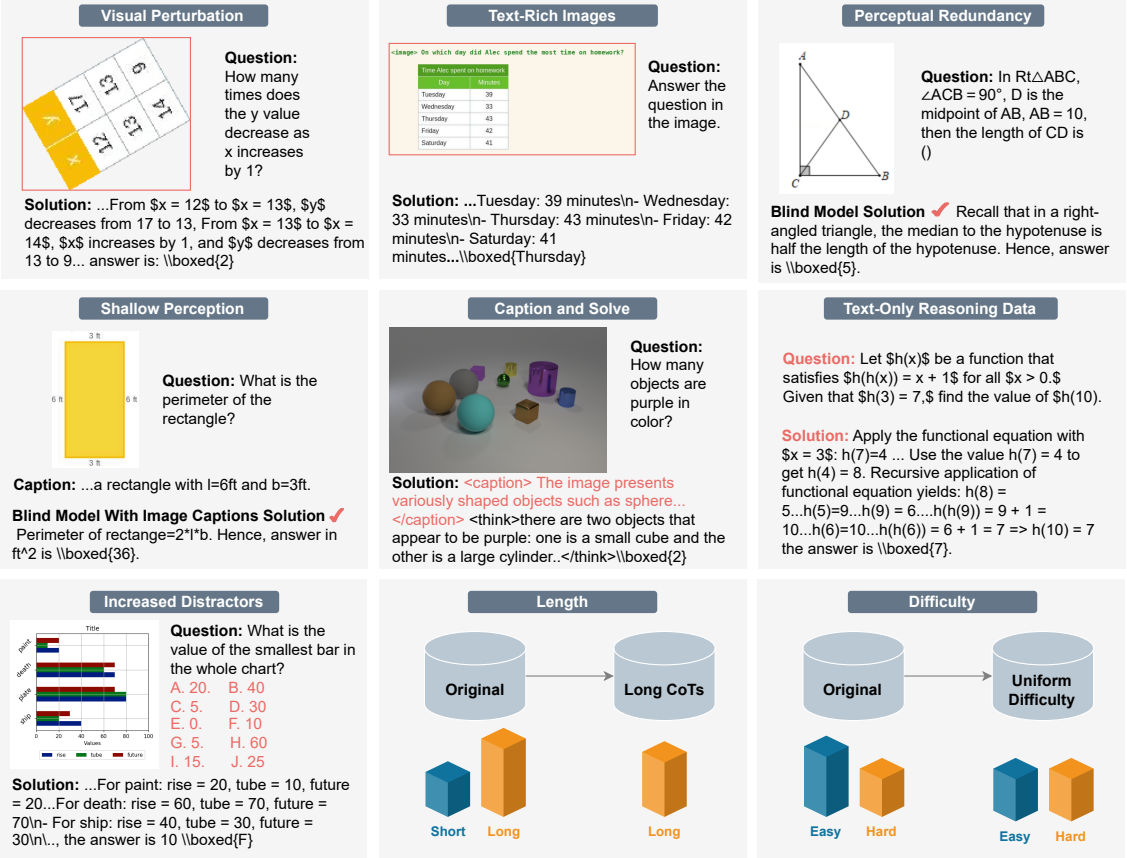


Figure 2. **Overview of data intervention strategies.** In this work, we curate a diverse set of data interventions to enhance the quality of VL reasoning CoT data. These methods target a range of skills, such as perception and problem-solving capabilities. We present details for each intervention in §3.2.

Problem-Solving Skill Enhancement. We briefly outline various data interventions for enhancing the *problem-solving* skills of the VL reasoner: (a) **text-only reasoning**, where we enhance problem-solving skills by exposing the VLM to novel problems and solutions from a high-performing text-only reasoning dataset [18]; (b) **increased distractors**, where we transform the original questions to have ten options and modify the original CoT accordingly to enhance the robustness of the VL reasoner; (c) **length**, where we split the original data into two halves based on the length of the CoT, and train on the longer half to encourage the VL reasoner to utilize additional inference-time compute by generating longer CoTs; and (d) **difficulty**, where we filter the dataset to ensure that the representation for each difficulty level is roughly balanced, increasing the overall challenge of our dataset (more details in App. 10).

3.3. Scaling Diverse Data Axes

There has been limited exploration of the impact of scaling data for VL reasoning. In addition, the introduction of a new modality (image) in VL space adds a new dimension to scaling training data. Hence, we study the impact of scaling diverse axes on the reasoning capabilities of the VLM: (a) **scaling images**, where we create three additional subsets of the original dataset, each of increasing size, resulting in four datasets in total: $[\mathcal{D}_{N/8}, \mathcal{D}_{N/4}, \mathcal{D}_{N/2}, \mathcal{D}_{N}]$.[4] (b) **scaling questions**, where we prompt the question generator to create new questions that are reasonable, solvable, and of similar difficulty to the original questions, conditioned on the image and question[59]; and (c) **scaling CoTs**, where we study the impact of increasing the number of CoT traces for a given image and question pair in the reasoning dataset. We present details about the scaling pipeline in App. 16.

4. Experimental Setup

Training Data. We use six source datasets: `ViRL` [64], `Math-LLaVA` [53], `R1-OneVision` [73], `ThinkLite-VL-Hard` [67], `LLaVA-CoT` [71], and `MMK12` [43]. We cap the number of instances at 50K to

[4]Here, $\mathcal{D}_{N/8}$ indicates that we randomly select $N/8$ instances from the original dataset. As we increase the data size, we are effectively scaling the number of unique images in the dataset.

focus on the impact of data quality in our data curation experiments. We then perform decontamination to remove any identical images from the evaluation datasets using exact deduplication with the pHash algorithm. We cap the number of instances at 50K to focus on the impact of data quality in our data curation experiments.

Generator Model. We fix the CoT generator model to a performant VLM, `Llama4-Scout` [44]. Specifically, it is an open-weights model consisting of 109B total parameters, of which 17B are active. This model enables efficient inference on a single A100 node using vLLM [62].

Models. We train the 3B and 8B Perception Language Models (PLMs) [12] for all data curation experiments. PLMs cannot generate vision-language CoTs, due to lack of appropriate instruction data. Thus, they serve as good base models that can be converted into VL reasoners with high-quality reasoning CoT data.

Evaluation. For our data curation experiments, we choose *five* VL reasoning downstream tasks as validation datasets for hill-climbing. These include `MathVerse` (testmini, vision-only subset) [76] containing 788 examples; `MathVista` (testmini) [39] containing 1000 examples; `MathVision` (testmini) [65] containing 304 examples; `MMMU-Pro` (vision) [74] containing 1730 examples, and `We-Math` (testmini) [50] consisting of 1740 examples focused on diverse knowledge granularity. Overall, we track the accuracy score averaged over these five evaluation datasets and two model trainings (PLM-3B and PLM-8B) during the data curation process. After the creation of our final HONEYBEE dataset, we also evaluate them on *five* more unseen evaluation datasets including `DynaMath` [80] and `LogicVista` [70]. Further, we evaluate the visual perception capabilities on `HallusionBench` [17]. In addition, we evaluate the general-purpose reasoning capabilities of a VLM on text-only reasoning datasets including math-centric `MATH500` [33] and science-centric `GPQA` [51].[5]

5. Experiments

5.1. Impact of Context Curation

Sourcing. We present results for training PLM-3B and PLM-8B on reasoning data constructed using diverse context sources in Table 2. Specifically, we compute the average performance across five VL reasoning datasets and rank the context sources from highest to lowest accuracy. Our experiments reveal that the choice of context source has a significant impact on downstream VL reasoning performance. Specifically, we observe a relative gap of 11.4 percentage points (pp) between the average performance of the lowest (`MMK12`, 36.0%) and highest performing source datasets (`ViRL`, 40.1%). This highlights the choice of context source has a significant impact on downstream VL reasoning performance.

Mixing. We next assess the impact of mixing contexts from different data sources. To this end, we combine examples from the previous stage for the top-2, top-4, and all source datasets, and select a random subset of size 50K. The results of training PLM-3B and PLM-8B on these mixtures are presented in Appendix Table 7. Interestingly, we find that mixing data from diverse sources does not improve the performance over the best-performing dataset, `ViRL`. This suggests that the heterogeneous distributions of these datasets may introduce conflicting biases, ultimately degrading the performance of mixed data.

5.2. Impact of Data Interventions

Starting from the best data from the previous stage, we note the impact of several data intervention experiments on VL reasoning capabilities in Appendix Table 3. Specifically, we target the strategies at improving the perception and problem-solving capabilities of the VLMs.[6]

We find that the best-performing data intervention for enhanced perception is the *caption and solve* strategy, which provides auxiliary visual signals about the image description in the reasoning CoT data (i.e., `<caption> description </caption> solution` $\boxed{answer}$). Specifically, this approach relatively improves the average accuracy by 3.3pp. In addition, we observe that the best-performing data intervention for enhanced problem-solving is the inclusion of *text-only reasoning* data in the training mixture. In particular, we observe a relative accuracy improvement of 7.5pp over the original dataset across multiple VL reasoning tasks and training runs. Further, we study the impact of text-only reasoning data on general-purpose reasoning datasets, `MATH500`, in Appendix 11.

5.3. Scaling to Million Samples

Scaling Trends Across Data Axes. To synthesize large-scale data, we study the scaling behavior of the best-performing dataset, `ViRL`, including the number of unique images, the number of questions per image, and the number of CoTs per (image, question) pair. We present the

[5] To ensure consistency, we use *accuracy* as the scoring metric and an identical prompt, instructing the model to always answer in `boxed` format, across all evaluations. Further, we use greedy decoding with maximum generation length of 2048 across all the evaluation datasets.

[6] We present the results for the best-performing configuration for the interventions. For instance, we experiment with diverse ways of implementing *caption and solve* method, and show the results for the best-performing variant.

	Average	MathVerse (testmini, vision-only)	MathVista (testmini)	MathVision (testmini)	MMMU-Pro (vision)	We-Math (testmini)	DynaMath*	LogicVista*	Hall. Bench* (image)	MATH500*‡	GPQA*‡ (diamond)
1B model scale											
PLM-1B [12]	25.9	17.8	48.6	15.1	15.8	35.5	30.3	23.0	50.3	15.2	7.1
InternVL-2.5-1B [11]	27.6	18.7	44.2	16.4	16.2	38.7	29.3	20.5	51.6	27.8	12.1
InternVL-3-1B-Instruct [78]	28.3	18.0	35.0	13.2	16.2	37.5	28.5	21.9	56.0	37.8	19.2
PLM-HoneyBee-1B	**36.2**	**29.4**	**53.7**	**23.0**	**18.8**	**50.6**	**39.3**	**28.6**	**56.1**	**36.8**	**25.8**
3B-4B model scale											
PLM-3B [12]	33.8	18.0	57.2	16.1	19.5	46.1	37.0	33.5	61.1	30.4	18.7
InternVL-2.5-4B [11]	41.5	28.7	**61.8**	24.7	**31.1**	55.6	40.7	32.1	65.5	49.4	25.3
Qwen2.5-VL-3B-Instruct [2]	42.6	35.0	58.9	23.7	29.8	49.2	42.5	**36.6**	**66.0**	62.0	22.7
PLM-HoneyBee-3B	**46.2**	**42.8**	61.2	**29.9**	28.4	**59.3**	**51.9**	**36.6**	65.0	59.4	**27.7**
7B-8B model scale											
PLM-8B [12]	34.6	19.3	59.3	17.1	20.5	47.9	36.7	30.8	64.0	34.0	16.2
InternVL-2.5-8B [11]	41.4	27.3	61.5	21.4	32.0	56.6	41.9	26.6	63.8	57.0	26.3
InternVL-3-8B-Instruct [78]	45.1	35.3	61.8	19.4	35.8	55.7	51.2	36.2	65.5	**69.6**	20.2
Qwen2.5-VL-7B-Instruct [2]	48.5	42.0	67.5	**27.6**	**37.1**	61.1	51.3	39.9	67.4	64.8	26.3
PLM-HoneyBee-8B	**49.8**	**43.0**	**68.2**	26.3	33.8	**66.1**	**53.3**	**41.3**	**68.8**	63.6	**33.3**

Table 1. **Performance of VL reasoners trained with HONEYBEE data.** We compare the accuracy of PLMs trained with the HONEYBEE data on diverse downstream evaluation datasets. We find that models trained on HONEYBEE achieve best-in-class performance across model sizes. Task-specific subsets or splits are indicated in brackets '()'. Datasets that were unseen during the data curation process are marked with *, and text-only reasoning datasets are marked with ‡.

Dataset	Avg.	PLM-3B	PLM-8B
ViRL	40.1	38.8	41.3
MathLLaVA	37.7	36.3	39.2
R1-Vision	37.3	35.7	38.8
ThinkLite	37.1	34.6	39.5
LLaVA-CoT	36.3	34.6	37.9
MMK12	36.0	34.6	37.3

Table 2. **Ranking the quality of the context from diverse datasets.** We train PLM-3B and PLM-8B on diverse source datasets. Then, we rank them in descending order (left to right) based on their average performance on VL reasoning downstream tasks. We present the detailed results in Appendix Table 6.

Method	Skill	Average
Original Data	-	40.1
Perception Enhancement		
Caption and Solve	Auxiliary Signal	41.4
Visual Perturb	Robustness	38.5
Text-Rich Images	Synthetic Images	38.8
Perceptual Redundancy	Feasibility wo/ image	36.5
Shallow Perception	Feasibility w/ caption	35.6
Problem-solving Enhancement		
Text-Only Data	Cross-modal transfer	43.1
Increased Distractors	Robustness	34.6
Length	Long Thinking	36.4
Uniform Difficulty	Hardness	34.6

Table 3. **Results for data intervention experiments.** We compare the performance of VL reasoners trained on datasets created using diverse data interventions. We find that augmenting the original CoT with image captions (*caption and solve*), and mixing with *text-only reasoning* data improves VL reasoning performance. The detailed results are presented in Appendix Table 8.

results for accuracy averaged across the validation downstream datasets in Figure 3. Interestingly, we find that the performance of the VL reasoners improves with scaling in images, new questions, and CoTs across model trainings of both PLM-3B and PLM-8B.

Putting Everything Together. Firstly, we include all the real data in the ViRL datasets, which consists 39K (image, question, and final answer) tuples. Since the amount of real data is limited, we scale the data by generating several CoTs for all (image, question) pairs, and synthetically creating new questions (Appendix Figure 6). Specifically, we generate 16 CoTs per real image and existing question pair (*scaling CoTs*). To retain the highest quality data, we filter out CoTs that do not lead to the correct final answer, leading to

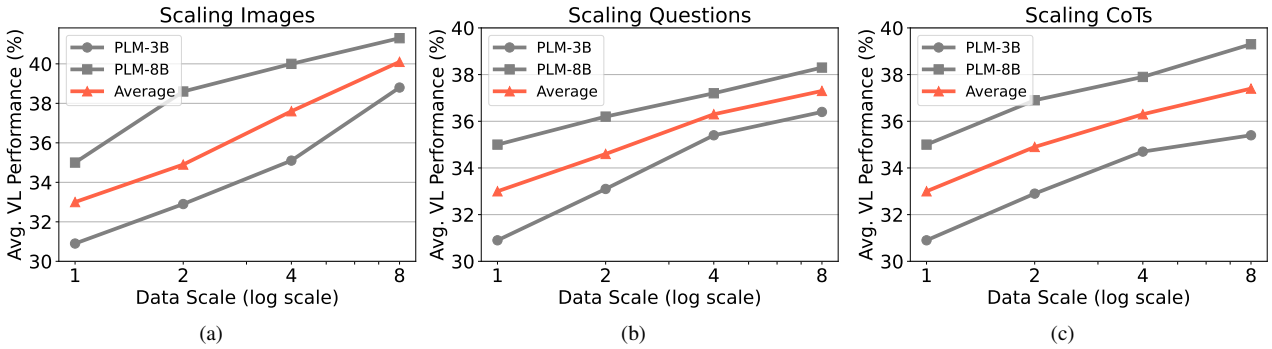


Figure 3. **Impact of scaling diverse data axes in VL reasoning data.** We train PLM-3B and PLM-8B on datasets of varying sizes (setup explained in §3.3). The results show that the reasoning performance consistently improves as we scale each data axis: (a) images, (b) synthetic questions per image, and (c) CoTs per (image, question) pair.

Statistic	Number (in K)
Total instances	2480
Number VL instances	1440
Number Text-Only instances	1040
Number Unique images	28
Number Unique questions	350
Avg. question length (words)	0.057
Avg. CoT length (words)	0.601

Table 4. **HONEYBEE statistics.**

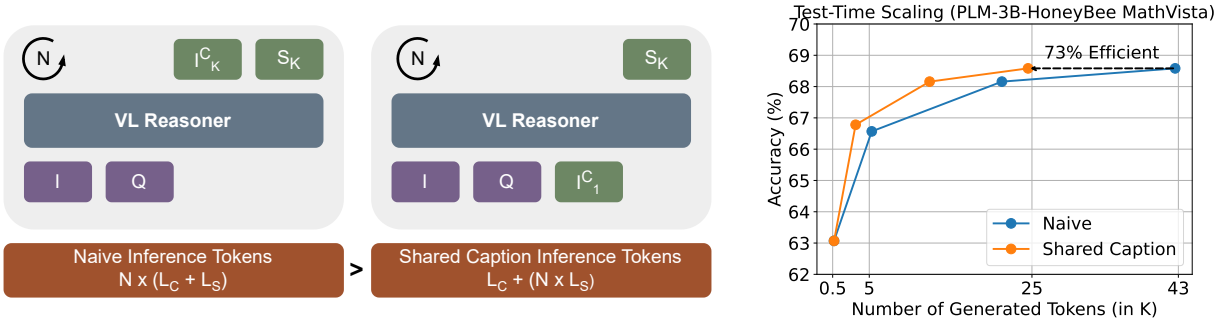
roughly 400K instances. To scale up the dataset, we generate 14 new questions per image, resulting in 15 questions per image in total (*scaling questions*). Since the final answers are unavailable for the newly generated questions, we instead produce 4 chains of thought (CoTs) per question and apply majority voting (an answer appearing three or more times) to approximate the final answer [49]. We then retain only those CoTs whose predicted answer matches this proxy final answer, yielding approximately 1M instances in the *scaling questions* setting. In parallel, we generate image captions for all images in the source datasets. Following the *caption and solve* strategy, we combine the (image, question, solution CoT) tuples from the scaling CoTs and questions pipelines with the image captions to construct (image, question, (image caption, solution CoT)) tuples. This process results in the construction of the VL subset of the HONEYBEE dataset, consisting of **1.5M** instances. We then merge this subset with the text-only reasoning data, consisting of **1M** instances, to obtain a high-quality, large-scale final HONEYBEE dataset of size **2.5M**.

Dataset Statistics. We present the dataset statistics in Figure 4. We highlight that HONEYBEE contains 28K unique images and 350K unique questions. The average length of CoTs (image caption and solution CoT combined) is approximately 600 words (780 tokens). We also highlight the occurrence of reasoning actions [14] in Appendix 13.

5.4. Training VL Reasoners with HONEYBEE

HONEYBEE elicits strong reasoning. Here, we assess the performance of PLMs of different sizes (1B, 3B, 8B) trained on the entire 2.5M HONEYBEE dataset. We also compare them against several high-performing VLMs capable of generating CoTs to solve VL reasoning tasks (Table 1). We find that VL reasoners trained with HONEYBEE achieve the highest average accuracy across all model size categories. In particular, PLM-HONEYBEE-1B achieves a relative performance improvement of 28pp over InternVL-3-1B-Instruct. Furthermore, PLM-HONEYBEE-3B and PLM-HONEYBEE-8B achieve relative gains of 8.4pp and 2.7pp over Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-8B-Instruct, respectively. Moreover, we perform a fine-grained comparison between the base PLM-3B and PLM-HONEYBEE-3B across diverse difficulty levels in the `MathVision` dataset (Figure 7). We find that HONEYBEE improves performance across all difficulty levels, reaching up to 100pp relative gains for level 2. This highlights that HONEYBEE can be used to enhance reasoning capabilities at various difficulty levels. Overall, we show that our high-quality curated data can outperform existing methods that provide little public information about their reasoning data.

HONEYBEE data scaling. We train PLM models on various subsets (50K, 250K, 2.5M) of the data. Results averaged across the five VL evaluation datasets are shown in Figure 1a. We observe that the performance of HONEYBEE-trained PLMs, across all sizes, continues to improve as the dataset size increases. In fact, we find that performance has not saturated even at the 2.5M scale. This suggests that, with sufficient training budget, one could further scale the size of HONEYBEE to achieve additional performance improvements. We show the scaling results across



(a) (b)

Figure 4. **Shared caption decoding for efficient test-time scaling (TTS).** (a) We illustrate the naive approach to TTS in VL reasoning and compare it to the proposed decoding strategy. (b) We present the accuracy trend as a function of the number of generated tokens (number of solution attempts up till 64) for PLM-3B trained with HONEYBEE on `MathVista` dataset.

individual datasets in Appendix Figure 9. Further, we compare the PLM-3B trained with the HONEYBEE data and existing CoT VL reasoning datasets in Figure 1b, and show massive relative gains upto 39pp across the five VL reasoning datasets. Apart from this, we use a small subset of our data to finetune the teacher model itself, reminiscent of self-improvement [54], and show that HONEYBEE data can be used to achieve reasoning performance from the generator model too (Appendix 18). We also perform a correlation analysis to examine the impact of our data design choices on the performance of the PLM models in Appendix 12.

RL training of HONEYBEE model. Supervised finetuning with CoT data serves as a warm start for RL training, enabling further improvements in reasoning capabilities [14, 35]. Thus, we perform RL training using the GRPO algorithm [52] on top of the PLM-HONEYBEE-3B model with VL verifiable data [36]. We present results on diverse VL reasoning tasks and compare them to those of a supervised and RL-tuned VL reasoner, OpenVLThinker-v1.2-3B, in Table 5. We find that RL training on top of HONEYBEE-SFT models outperforms OpenVLThinker-v1.2-3B, with a 9.2pp relative gain in average accuracy. We present the detailed results in Appendix Table 10.

Method	Average
OpenVLThinker-v1.2-3B [14]	42.3
PLM-HONEYBEE-3B	44.3
PLM-HONEYBEE-3B-GRPO	**46.2**

Table 5. **RL training on top of HONEYBEE trained VLM.** We present results for training PLM-3B with HONEYBEE data using supervised finetuning, followed by a round of RL training.

5.5. Efficient test-time scaling with Shared captions

We observe that each CoT in the HONEYBEE dataset has two parts: an understanding component (I^C, image caption tokens) and a problem-solving component (S, solution tokens), so $C = [I^C; S]$. In test-time scaling (TTS) methods like self-consistency [66], we generate $N > 1$ CoTs for a reasoning problem (I, Q) and use majority voting over answers. The naive TTS approach generates the full CoT N times. Instead, we propose *shared captions*: generate the full CoT once as (I^C_1, S_1), then reuse I^C_1 as context for subsequent solution generations S_K (Figure 4a). This reduces number of generated tokens and, since inference FLOPs scale with token count [24], improves inference efficiency. We empirically test this with PLM-3B trained on HONEYBEE using `MathVista` (Figure 4b), generating $N = 64$ solutions at temperature 0.7 per (image, question) pair. The naive approach produces 42.6K tokens (671 per attempt), while *shared captions* achieves similar performance with only 24.5K tokens (280 captioning, 390 problem-solving per attempt), a 73% reduction in tokens and FLOPs. Thus, HONEYBEE enables efficient and strong VL reasoning.

6. Conclusion

We present HONEYBEE, a high-quality and large-scale VL reasoning dataset. Future work can assess the impact of our insights on general-purpose data curation for VL training, particularly for skills that go beyond reasoning, such as VQA. Moreover, we have focused only on data curation for single images, but it would be pertinent to extend this approach to reasoning over multiple images. Our work lays a strong foundation for data research in VL reasoning.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida,

Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 1

[3] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023. 3

[4] Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q Tran, and Mehran Kazemi. Smaller, weaker, yet better: Training llm reasoners via compute-optimal sampling. *arXiv preprint arXiv:2408.16737*, 2024. 3

[5] Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil T Heffernan, and Kyle Lo. Drawedumath: Evaluating vision language models with expert-annotated students' hand-drawn math images. *arXiv preprint arXiv:2501.14877*, 2025. 1

[6] Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, Eric Chung, and Chris Alexiuk. Llama-nemotron: Efficient reasoning models, 2025. 1, 3

[7] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 11

[8] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023. 3

[9] Liang Chen, Hongcheng Gao, Tianyu Liu, Zhiqi Huang, Flood Sung, Xinyu Zhou, Yuxin Wu, and Baobao Chang. G1: Bootstrapping perception and reasoning abilities of vision-language model via reinforcement learning. *arXiv preprint arXiv:2505.13426*, 2025. 1

[10] Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025. 1

[11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6, 1

[12] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. 5, 6, 11

[13] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[14] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *arXiv preprint arXiv:2503.17352*, 2025. 7, 8, 11

[15] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025. 2

[16] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 3

[17] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14375–14385, 2024. 5

[18] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025. 1, 3, 4, 2, 7, 10

[19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 11

[20] Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025. 1, 4

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[23] Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhu Chen. Visualwebinstruct: Scaling up multimodal instruction data through web search. *arXiv preprint arXiv:2503.10582*, 2025. 2

[24] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 8

[25] Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, et al. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*, 2025. 3

[26] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 3

[27] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 1

[28] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. 2

[29] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities. *arXiv preprint arXiv:2403.04706*, 2024. 1

[30] Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13 (9):9, 2024. 3, 1

[31] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Kamal Mohamed Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Joshua P Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander T Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alex Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-LM: In search of the next generation of training sets for language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2

[32] Yuting Li, Lai Wei, Kaipeng Zheng, Jingyuan Huang, Linghe Kong, Lichao Sun, and Weiran Huang. Vision matters: Simple visual perturbations can boost multimodal math reasoning. *arXiv preprint arXiv:2506.09736*, 2025. 1, 3, 2

[33] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. 5

[34] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 2

[35] Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy. *arXiv preprint arXiv:2506.13284*, 2025. 1, 8

[36] lmmslab. lmms-lab/multimodal-open-r1-8k-verified · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/lmms-lab/multimodal-open-r1-8k-verified, 2025. 8

[37] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 2

[38] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 2

[39] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel

Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1, 3, 5

[40] Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, et al. Datadecide: How to predict best pretraining data with small experiments. *arXiv preprint arXiv:2504.11393*, 2025. 4

[41] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*, 2024. 3

[42] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1

[43] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 4, 5

[44] AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4 (7):2025, 2025. 1, 5

[45] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*, 2024. 1

[46] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. 1

[47] OpenAI. Openai o3 and o4-mini system card. 1

[48] OpenAI. Gpt-4v(ision) system card. 2023. 1

[49] Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason Weston, and Jane Yu. Self-consistency preference optimization. *arXiv preprint arXiv:2411.04109*, 2024. 7

[50] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024. 5

[51] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. 5

[52] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 8

[53] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 1, 3, 4, 2, 5

[54] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023. 3, 8

[55] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[56] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[57] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 3, 1

[58] NovaSky Team. Sky-t1: Train your own o1 preview model within 450 usd. https://novasky-ai.github.io/posts/sky-t1, 2025. 1

[59] Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024. 4, 1, 7

[60] Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *Advances in Neural Information Processing Systems*, 37:34737–34774, 2024. 1

[61] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024. 1

[62] The vLLM Team. Llama 4 in vLLM — blog.vllm.ai. https://blog.vllm.ai/2025/04/05/llama4.html, 2025. 5

[63] Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. *arXiv preprint arXiv:2401.13311*, 2024. 3, 2

[64] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025. 4, 2, 5

[65] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 5, 1

[66] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 8, 1

[67] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025. 4, 5

[68] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024. 3

[69] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1, 3

[70] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. 5

[71] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 1, 4, 2, 5

[72] Yuqing Yang, Yan Ma, and Pengfei Liu. Weak-to-strong reasoning. *arXiv preprint arXiv:2407.13647*, 2024. 3

[73] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025. 4, 1, 2, 5

[74] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 5, 1, 2, 3

[75] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022. 1

[76] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 5, 1, 3

[77] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*, 2024. 3, 1, 2

[78] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6

[79] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023. 2

[80] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024. 5