Phi-4-reasoning-vision-15B Technical Report

Jyoti Aneja, Michael Harrison, Neel Joshi,
Tyler LaBonte, John Langford, Eduardo Salinas

Microsoft Research

https://aka.ms/Phi-4-reasoning-vision https://huggingface.co/microsoft/Phi-4-reasoning-vision-15B
https://aka.ms/Phi-4-r-v-FoundryLabs https://github.com/microsoft/Phi-4-reasoning-vision-15B

We present Phi-4-reasoning-vision-15B, a compact open-weight multimodal reasoning model, and share the motivations, design choices, experiments, and learnings that informed its development. Our goal is to contribute practical insight to the research community on building smaller, efficient multimodal reasoning models and to share the result of these learnings as an open-weight model that is good at common vision and language tasks and excels at scientific and mathematical reasoning and understanding user interfaces. Our contributions include demonstrating that careful architecture choices and rigorous data curation enable smaller, open-weight multimodal models to achieve competitive performance with significantly less training and inference-time compute and tokens. The most substantial improvements come from systematic filtering, error correction, and synthetic augmentation—reinforcing that data quality remains the primary lever for model performance. Systematic ablations show that high-resolution, dynamic-resolution encoders yield consistent improvements, as accurate perception is a prerequisite for high-quality reasoning. Finally, a hybrid mix of reasoning and non-reasoning data with explicit mode tokens allows a single model to deliver fast direct answers for simpler tasks and chain-of-thought reasoning for complex problems.

1 Introduction

Phi-4-reasoning-vision-15B is a compact open-weight multimodal reasoning model that balances reasoning power, efficiency, and training data needs. It is a broadly capable model that allows for natural interaction for a wide array of vision-language tasks and excels at math and science reasoning and understanding user interfaces, as shown in Figure 1. Beyond these general capabilities, our model presents an appealing value relative to current open-weight models, pushing the Pareto frontier of the trade-off between accuracy and compute costs. We achieve competitive accuracy with much slower models that require ten times or more compute time and tokens, and better accuracy than similarly fast models, particularly when it comes to math and science reasoning, as shown in Figure 2.

In this report, we share the motivations, design choices, experiments, and learnings that informed its development, as well as an evaluation of the model's performance and guidance on how to use it. Our goal is to contribute practical insight to the community on building smaller, efficient multimodal reasoning models and to share an open-weight model that is competitive with models of similar size at general vision-language tasks, excels at computer use, and at scientific and mathematical multimodal reasoning.

1.1 Focus on Smaller and Faster Vision–Language Models

Many popular vision–language models (VLMs) have trended towards growing in parameter count and the number of tokens they consume and generate. This leads to increased training and inference-time cost and latency, impeding their usability for downstream deployment, especially in resource-constrained or interactive settings.

A growing countertrend towards smaller models aims to boost efficiency, enabled by careful model design and data curation—a goal pioneered by the Phi (Gunasekar et al., 2023) family of models and furthered by Phi-4-reasoning-vision-15B. We specifically build on learnings from the Phi-4 (Abdin et al., 2024) and Phi-4-Reasoning (Abdin et al., 2025) language models and show how a multimodal model can be trained

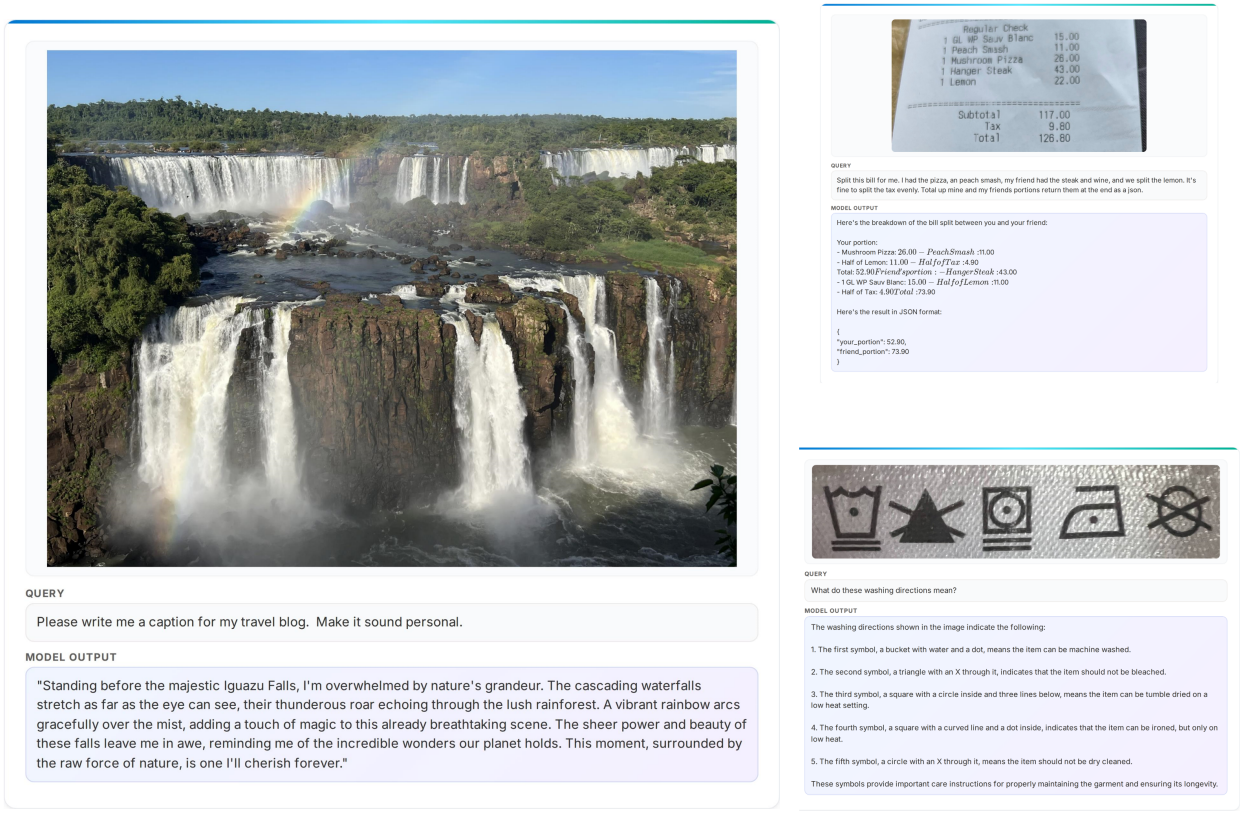


Figure 1: Phi-4-reasoning-vision-15B can help with a wide range of everyday tasks, from writing travel captions and interpreting receipts to reading garment care instructions.

to cover a wide range of vision and language tasks without relying on extremely large training datasets, architectures, or excessive inference-time token generation. Our model is intended to be lightweight enough to run on modest hardware while remaining capable of structured reasoning when it is beneficial. Our model was trained with far less compute than many recent open-weight VLMs of similar size. We used just 200 billion tokens of multimodal data leveraging Phi-4-Reasoning (trained with 16 billion tokens) based on a core model Phi-4 (400 billion unique tokens), compared to more than 1 trillion tokens used for training multimodal models like Qwen 3 VL (Bai et al., 2025), Kimi-VL (Team et al., 2025b), and Gemma3 (Team et al., 2025a). We therefore present a compelling option compared to existing models, pushing the Pareto frontier of the trade-off between accuracy and compute costs.

2 Architecture and Training

Training a multimodal reasoning model raises numerous questions and design decisions around model architecture, dataset quality and composition, training curriculum, and the interaction between reasoning-heavy and non-reasoning perception-focused tasks. All of these choices affect the learned behavior.

2.1 Early vs. Mid Fusion

There are several options for model architecture based on when and how visual and textual information is fused. In late or mid-fusion models, a vision encoder first converts images into a compact set of visual tokens via a pretrained image encoder, which are then projected into the language embedding space and injected into a pretrained LLM (Liu et al., 2023). This approach enables meaningful cross-modal reasoning while preserving the strengths and scalability of large unimodal models. This approach keeps training and

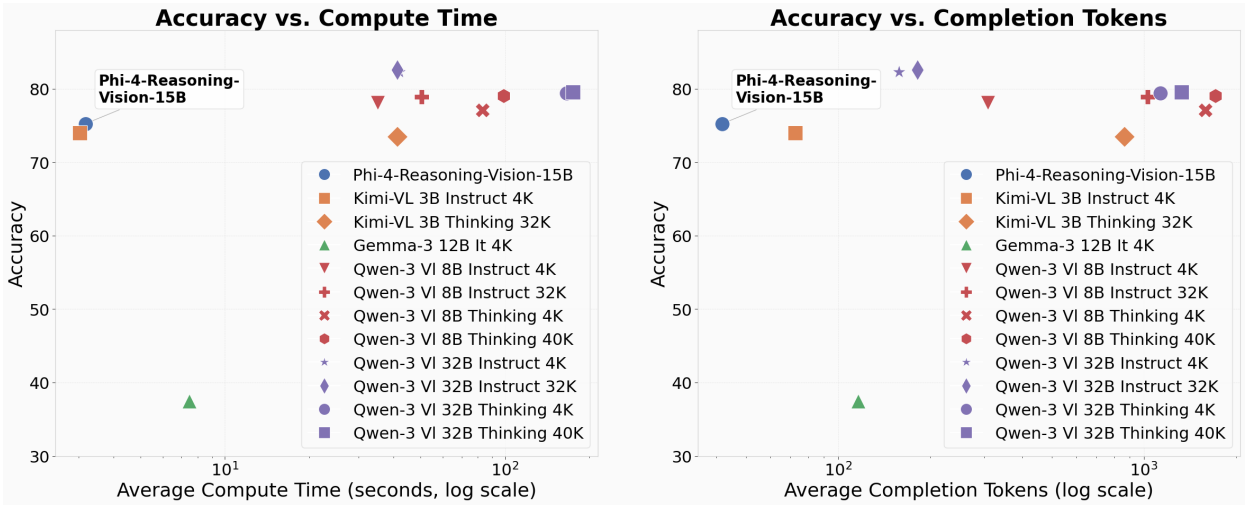


Figure 2: Phi-4-reasoning-vision-15B presents a compelling option compared to existing models, pushing the Pareto frontier of the trade-off between accuracy and compute costs. We achieve competitive performance with much slower models that require more time and tokens, and higher accuracy than similarly fast models. These values were computed by averaging accuracy, time, and output token counts for a subset of 4 benchmarks: $\text{ChartQA}_{\text{TEST}}$, $\text{MathVista}_{\text{MINI}}$, MMMU_{VAL}, and ScreenSpot_v2.

inference costs manageable, as it can utilize the power of pretrained components that have typically been trained on trillions of tokens.

Early-fusion models, by contrast, process all image patches and text tokens into a single transformer, allowing unrestricted cross-attention across modalities throughout the network (Team, 2025). While this can yield richer joint representations and tighter visual–textual grounding, it significantly increases compute, memory, and data requirements. Given our goal of creating a highly performant model with less compute and data, we use a mid-fusion architecture. It offers a practical trade-off between expressivity and efficiency without the heavy cost of full early fusion.

2.2 Vision Encoder and Image Processing

We build on the SigLIP-2 vision encoder (Tschannen et al., 2025) and the Phi-4-Reasoning backbone, as shown in Figure 3. In our previous work, we found that multimodal language models sometimes struggled to solve tasks, not because of a lack of reasoning proficiency, but rather because of an inability to identify and extract relevant perceptual information from the image (Balachandran et al., 2024). This problem compounds when considering computer-use and multimodal grounding tasks. In particular, desktop screens and browsers are information-dense with relatively small interactive elements, making fine-grained high-resolution feature extraction a prerequisite for agentic applications.

With high-resolution multimodal benchmarks increasing in relevance, several open-source multimodal language models have adapted their methodologies accordingly, e.g., Gemma3 (Team et al., 2025a) uses pan-and-scan, NVILA (Liu et al., 2025b) uses dynamic S^2, and Qwen3-VL (Bai et al., 2025) uses a bespoke vision encoder which operates at native resolution. However, their trade-offs are difficult to understand across different datasets and hyperparameters. To explore these options, we conducted a large-scale ablation of several vision encoder and image processing techniques, with the goal of understanding and maximizing grounding performance.

We trained a smaller (5B) variation of our model on a dataset of 10M image-text pairs, primarily composed of computer-use and GUI grounding data and experimented with several vision encoder configurations:

- **Dynamic S^2** (Liu et al., 2025a): similar to S^2, but resizes to a rectangular resolution chosen to minimize distortion while admitting a tiling by 384×384 squares.

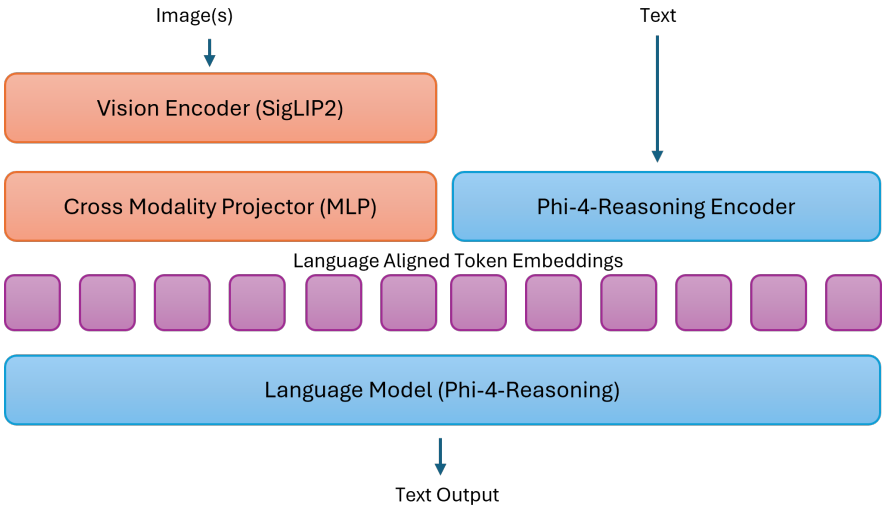


Figure 3: Overview of the Phi-4-reasoning-vision-15B mid-fusion architecture. Images are processed by a SigLIP-2 vision encoder and projected into the language embedding space via a cross-modality projector (MLP). The resulting visual "soft" tokens are interleaved with text tokens and fed into the Phi-4-Reasoning language model.

- **Multi-crop**: crops the image into (potentially overlapping) 384×384 squares; sends each through the vision encoder and concatenates features on the token dimension.
- **Multi-crop with S^2**: similar to multi-crop but uses S^2 to broaden the receptive field, i.e., crops the image into (potentially overlapping) 1536×1536 squares, performs S^2, and concatenates features on the token dimension.
- **Dynamic resolution**: a natively dynamic resolution vision encoder; we primarily used the NaFlex variant of the SigLIP-2 encoder (Tschannen et al., 2025) and adjusted the minimum and maximum number of patches.

Our primary finding is that dynamic resolution vision encoders with a large number of visual tokens perform uniformly well, and the best on high-resolution datasets. It is particularly interesting to compare dynamic resolution with 2048 vs. 3600 maximum tokens: the latter roughly corresponds to native HD 720p resolution and enjoys a substantial boost on high-resolution benchmarks, particularly ScreenSpot-Pro. Reinforcing the high-resolution trend, we find that multi-crop with S^2 outperforms standard multi-crop despite using fewer visual tokens (i.e., fewer crops overall). Finally, it is worth noting that the dynamic resolution technique produces the most tokens on average; due to their tiling subroutine, S^2-based methods are constrained by the original image resolution and often only use about half the maximum tokens.

Open research questions. While an increase in the resolution of the vision encoder substantially improves performance on high-resolution reasoning tasks, it comes at the cost of efficiency due to the quadratic

Table 1: Results with different resolution handling approaches on MathVista (Lu et al., 2024), ScreenSpot (Cheng et al., 2024), ScreenSpot-Pro (Li et al., 2025), and V*Bench (Wu and Xie, 2023). We have **bolded** the top two configurations on each benchmark. These experiments are done on a smaller, 5B, variation of our model created for testing purposes.

Method	Max Tokens	MathVista	ScreenSpot	ScreenSpot-Pro	V*Bench
Dynamic-S^2	3096	42.9	78.4	9.4	52.9
Multi-crop	3096	43.4	67.8	5.4	51.8
Multi-crop with S^2	2048	43.4	**79.1**	**10.6**	**57.1**
Dynamic resolution	2048	**45.2**	**81.5**	9.2	51.3
Dynamic resolution	3600	**44.9**	79.7	**17.5**	**56.0**

Table 2: Training recipe for Phi-4-reasoning-vision-15B. Trainable modules are indicated with ✓; frozen modules with ×.

Stage	MLP	Vision Encoder	LLM	Data
1. MLP Pretraining	✓	×	×	Image–text alignment
2. Instruction Tuning	✓	✓	✓	Single-image instruction tuning
3. Long Context & RAI	✓	✓	✓	Long content, multi-image, RAI

complexity of attention with respect to the context length. With that said, each featurization technique we tested operates independently of the text prompt. It is an open question how to leverage text-conditioning to most efficiently tokenize the image—for example, if a specific question is asked about a high-resolution scene, the background could be encoded in a lower resolution to save on tokens. Similar ideas are present in the literature (e.g., the Q-Former from BLIP-2 (Li et al., 2023)), but their initial promise has not yet been proven out for agentic tasks.

2.3 Training Recipe

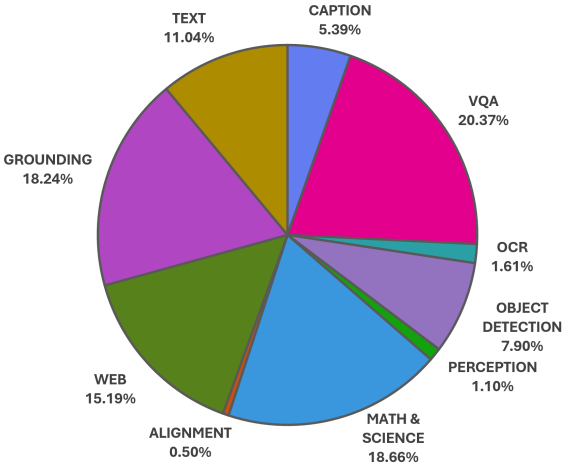
Phi-4-reasoning-vision-15B was trained in three stages, summarized in Table 2. The 1st stage trains the MLP only, with the rest of the model frozen, to warm-up the MLP from its random initialization. This stage is relatively light and uses a small amount of clean image-captioning data to create initial alignment between the vision-encoder and LLM backbone. We experimented with larger amounts of pretraining in this initial stage and saw no benefit. Stage 2 is the bulk of the training and trains the whole model unfrozen on a larger amount of single-image visual instruction tuning data. Stage 3 is a medium sized stage that focuses on long-context, multi-image, and safety (RAI) training. Table 3 lists the key optimization hyperparameters and training sample and token counts for each stage. All stages use the AdamW optimizer, bf16 mixed precision, DeepSpeed ZeRO-1, and train for one epoch. More details on each stage:

Stage 1: MLP Pretraining. Only the cross-modality projector (MLP) is trained while the vision encoder and language model remain frozen. This stage aligns the visual feature space of SigLIP-2 with the text embedding space of Phi-4-Reasoning, establishing a shared representation before any other parameters are updated.

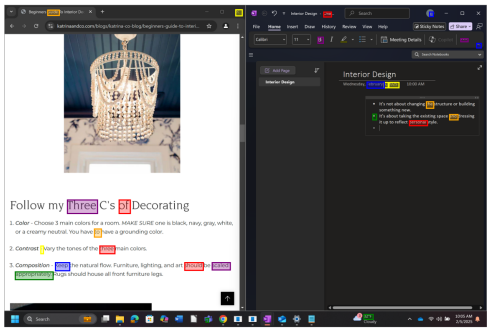
Stage 2: Instruction Tuning. All model components—the MLP, vision encoder, and language model—are jointly trained on single-image instruction-tuning data. This stage constitutes the bulk of training and covers the full range of tasks: visual question answering, mathematical and scientific reasoning, grounding, captioning, OCR, and computer-use. The mixture includes both reasoning traces (with `<think>` tokens) and direct-response samples (with `<nothink>` tokens) as described in Section 4.

Table 3: Training hyperparameters by stage.

Hyperparameter	Stage 1	Stage 2	Stage 3
Learning rate	1×10^{-3}	2×10^{-5}	7×10^{-7}
LR schedule	Cosine	Cosine w/ min LR	Cosine w/ min LR
Min LR ratio	–	0.1	0.1
Warmup	3% of steps	500 steps	500 steps
Weight decay	0	10^{-4}	10^{-4}
Adam (β_1, β_2)	(0.9, 0.999)	(0.9, 0.95)	(0.9, 0.95)
Max grad norm	1.0	1.0	1.0
Global batch size	1024	1920	1920
Max sequence length	2048	8192	16384
Training samples	2.0M	62.8M	3.2M
Training tokens	1.4B	188.5B	12B



Training Data Composition by Category



Web/Grounding

{"image":"web_41352592_scroll_2/screenshot.png", "conversations":
[{"from": "human", "value": "where's the following UI element? Give me a relative bounding box The text 'Competing Interests' is located within the table of contents on the right side of the webpage, below the 'Table of Contents' heading.\nThe text 'Competing Interests' is displayed in a standard font, grey color, and is part of the list of items in the table of contents.\n<image>"},
{"from": "gpt", "value": "**<nothink>**[0.5830, 0.1090, 0.6250, 0.1180]"}]}

{"image": "MatplotlibChartPipeline_ridgeline_1-286.png", "conversations":
[{"from": "human", "value": "Which city had the highest expatriate population in 2022?"},
{"from": "gpt", "value": "**<think>**To find the city with the highest expatriate population in 2022, we look at the topmost line in the Ridgeline plot for the year 2022. Dubai is at the top with the highest value.**</think>**Dubai"}

Training with Mixed Non-Reasoning/Reasoning Data

Original caption: Sunset Weddings Cabo

Our Short Recap: A bride stands on a beach with her arms outstretched, showcasing a long veil that billows behind her.

Our Medium Recap: The image captures an elegant bride on a sandy beach, facing the ocean with her back to the camera. Her arms are wide open, extending a sheer, flowing veil that stretches across the frame. The scene is serene and picturesque, with the waves gently crashing on the shore and a clear, vibrant sky above. Rocky formations are visible to the right, highlighting the natural beauty surrounding her.

GPT-4o Captions of images

Figure 4: Training data composition and examples for the Stage 2 training of Phi-4-reasoning-vision-15B. The Stage 3 data is designed to have a similar composition.

Stage 3: Long Context, Multi-Image, and RAI. The full model continues training on specialized data: long-document understanding, multi-image and sequential-image tasks, and additional responsible AI (RAI) data. This stage extends the model's capabilities to handle longer contexts and multi-turn visual interactions while reinforcing safety alignment.

Table 7 in the Appendix lists the public training data sources used across all stages, grouped by category. The majority of our data originates from public open-source datasets which were filtered and improved as described in Section 3.

3 Training Data

As with its language backbone Phi-4-Reasoning (Abdin et al., 2025), Phi-4-reasoning-vision-15B was trained with a deliberate focus on data quality. Our final data mix consists of data primarily from three sources: open-source vision-language datasets which were meticulously filtered and improved, high-quality domain-specific data from other Microsoft teams, and high-quality data from targeted acquisitions. The overwhelming majority of our data lies in the first category: data which originated as open-source data, after which a significant amount of effort was dedicated to filtering and improving, whether by removing low-quality datasets or records, programmatically fixing errors in data formatting, or using open-source images as seeds to synthetically generate higher-quality accompanying text.

3.1 Data Quality

The process of improving open-source data began by simply spending time manually sifting through data. Going through samples from each dataset, we found that 5–10 minutes per dataset was enough to classify in

one or more of the following categories:

- **Excellent-quality data**: the text components of the data consist of high-quality questions paired with correct answers. The threshold for "excellent" data is somewhat category dependent; for example, high-quality caption data might look different from high-quality chart QA data.
- **Good questions with wrong answers**: the text components of the data consist of high-quality questions, answerable from the image, with some portion of incorrect answers. This category arises most commonly with diagram/math/STEM QA.
- **Low-quality questions**: the text components of the data contain some number of low-quality questions, which are either nonsensical or not answerable from the given image.
- **Low-quality images**: the images themselves are too repetitive or have fundamental errors (for example, a synthetic dataset of LaTeX diagrams where text and figures tend to overlap chaotically).
- **High-quality with formatting errors**: the text component of the data contains formatting errors for many records, probably introduced during some processing stage; for example: all answers in a different format than what the prompt requests, misspelled image tags, final answers contained within reasoning blocks, etc.

Excellent-quality data was kept mostly unchanged, except perhaps for minor formatting improvements. For data with wrong answers or poor-quality captions, we re-generated answers or captions using GPT-4o and o4-mini, and, where appropriate, used the same models in verification or majority-voting pipelines. Not all such attempts succeeded, and we excluded a number of datasets with high percentages of wrong answers. We made some attempts to improve the data with low-quality questions, but we did not find much success with a naive approach (asking models for high-quality questions to complement images) except in some very special cases. However, when the images themselves were high quality, we used these as seeds to generate caption data or simple VQA, with the questions perhaps of a different flavor than the original text. We excluded datasets where the images themselves were fundamentally flawed. We fixed many formatting or logical errors, of which we found a surprising number across open-source datasets.

We employed a variety of techniques to get more mileage out of datasets, often with basic reformatting or diversification techniques, or using images as seeds to generate new image-text pairs. Some examples are as follows:

1. Based on the belief that Phi-4-Reasoning, as a language model backbone, can solve many VQA math problems provided that it can adequately interpret the mathematical elements of an image, we took every image from math/science/logic datasets and generated a detailed description of the image. This means that for all such domain-specific data, our data mix contains multiple records with the same image: one with the original QA and one with a caption-style description.
2. Due to the limited amount of high-quality training data, we often asked our data to perform double-duty; for example, instead of having instruction-following data separate from the domain-specific data, we modified the text portion of data with ground-truth QA pairs to request and provide the answer in a specific format.
3. After generating high-quality caption data from open-source images, we created multi-image data by creating records in scrambled and caption-matching formats. For the former, $\sim$5 images are given, and then captions are requested in a random order, and occasionally additional images are sprinkled in later. For the latter, $\sim$5 images are given, and the request is to match captions to images. We believe that such data improves the model's ability to attend to correct images in certain multi-image scenarios.
4. With a similar goal to item 3, we generated "what's changed?" style data from pairs or triples of sequential screenshots, with the belief that such data improves the model's ability to better navigate images in real-time, as is necessary for CUA or robotics scenarios.
5. We found that some datasets use overly complicated or over-engineered prompts (that is, the user half of the text portion) in their VQA data, which is likely to teach the model to only succeed in answering perfectly structured questions. We use a variety of human prompts to teach more robustness to the model.

Table 4: Varying the ratios of math and CUA data. Increasing math data by 3× while keeping computer-use data constant improves both math and computer-use benchmarks.

General	Math	CUA	Total	MMMU-CoT	MathVista	ScreenSpot-V2
1M	150K	450K	1.6M	44.0	37.4	48.2
1M	150K	850K	2.0M	44.1	37.3	60.0
1M	450K	450K	1.9M	45.3	36.0	48.3
1M	450K	850K	2.3M	43.4	38.9	**63.1**
1M	150K	150K	1.3M	44.2	36.9	29.8
1M	150K	250K	1.4M	**45.4**	37.4	37.7

To supplement the improved open-source data, we utilize high-quality datasets shared with us by several teams at Microsoft, as well as several math-specific datasets which were acquired during training of the Phi-4 language model, and also some domain-specific curated data; for example, LaTeX-OCR data generated by processing and rendering equations from arXiv documents.

Coordinate normalization. Note that all spatial coordinates in our data are normalized to the range $[0.0, 1.0]$ relative to the image dimensions. This ensures a consistent representation across images of varying resolutions, and thus our model also always outputs normalized coordinates as well.

3.2 Mathematics and Science vs. Computer-Use Data Proportion

One of our goals was to train a model which was simultaneously proficient at both mathematics and computer-use. It is an open question in the research community to understand how datasets should be structured to induce generalizable representations across diverse reasoning tasks. Importantly, how data scale affects reasoning performance can lead to starkly different design decisions, e.g., training a single model on a large dataset vs. multiple models with targeted post-training for mathematics and computer use.

Research on long-tailed classification robustness has suggested that balancing the data, or removing data from overrepresented tasks or subgroups, is an effective method for ensuring uniformly good performance (Buda et al., 2018; Idrissi et al., 2022; Chaudhuri et al., 2023). Nevertheless, these insights are at odds with the scale-is-all-you-need data paradigm. We conducted a suite of experiments to better understand optimal data scale and ratios for multimodal reasoning tasks of math and science reasoning vs. computer use – our key focus areas for the model.

We trained a smaller variation of our model (5B parameters), while varying the amount of mathematics and computer-use data for each run. Each dataset included the same subset of 1M general image-text pairs as a baseline. For mathematics data, we used the same dataset of 150K multimodal records, optionally duplicating each one 3 times. Next, we included up to 450K computer-use records, and optionally an additional 400K from Phi-Ground (Zhang et al., 2025).

Our finding is that it appears possible for a single model to have uniformly superior performance across multiple reasoning domains. In general, multimodal mathematics performance was not harmed by additional computer-use data, and vice versa. The most impressive improvements were obtained on ScreenSpot-V2 by including the Phi-Ground dataset; its high specialization to GUI grounding reinforces the importance of targeted novel data collection. It is also worth noting that increasing mathematics data while keeping computer-use data constant still improves computer-use benchmarks.

Open research questions. Our experiments were conducted at a fairly small data scale wherein the model has not yet become saturated; in particular, overall performance correlated well with total data. A clear open question is to study the effects of data proportion at a scale which challenges the edge of current models' capabilities: do our insights about strong uniform performance hold, or do trade-offs between different reasoning tasks become more obvious at larger scales? Moreover, our imbalance ratios were fairly mild, with a 7.5% ratio of mathematics data to total data at the worst. While well-studied in traditional machine learning

settings such as long-tailed classification, understanding data dynamics at more extreme ratios (1% or less) is an open problem, especially for performance on competing reasoning tasks.

4 Mixed Non-Reasoning and Reasoning

In language-only settings, reasoning traces have improved performance on many tasks, but they require additional compute which adds undesired latency. In multimodal settings, this tradeoff is less clear-cut: for tasks such as image captioning and optical character recognition (OCR), reasoning is often unnecessary and can even be harmful, while mathematical and scientific problem-solving benefit from multi-step reasoning. Thus, the choice of when to reason or not can be quite nuanced.

4.1 Training Approaches for Multimodal Reasoning Models

Language-only reasoning models are typically created through supervised fine-tuning (SFT) or reinforcement learning (RL): SFT is simpler but requires large amounts of expensive reasoning trace data, while RL reduces data requirements at the cost of significantly increased training complexity and compute. Multimodal reasoning models follow a similar process, but the design space is more complex. With a mid-fusion architecture, the first decision is whether the base language model is itself a reasoning or non-reasoning model. This leads to several possible training pipelines:

1. **Non-Reasoning LLM → Reasoning Multimodal Training:** Reasoning and multimodal capabilities are trained together.
2. **Non-Reasoning LLM → Non-Reasoning Multimodal → Reasoning Multimodal Training:** Multimodal capabilities are learned first, then reasoning is added.
3. **Reasoning LLM → Reasoning Multimodal Training:** A reasoning base is used, but all multimodal data must include reasoning traces.
4. **Our approach: Reasoning LLM → Mixed Non-Reasoning / Reasoning Multimodal Training.** A reasoning-capable base is trained on a hybrid data mixture, learning when to reason and when to respond directly.

Approaches 1 and 2 offer flexibility in designing multimodal reasoning behavior from scratch using widely available non-reasoning LLM checkpoints but place a heavy burden on multimodal training. Approach 1 must teach visual understanding and reasoning simultaneously and requires a large amount of multimodal reasoning data, while Approach 2 can be trained with less reasoning data but risks catastrophic forgetting, as reasoning training may degrade previously learned visual capabilities. Both risk weaker reasoning than starting from a reasoning-capable base. Approach 3 inherits strong reasoning foundations, but like Approach 1, it requires reasoning traces for all training data and produces reasoning traces for all queries, even when not beneficial.

4.2 Our Approach: A Mixed Reasoning and Non-Reasoning Model

Phi-4-reasoning-vision-15B adopts the 4th approach listed previously, as it balances reasoning capability, inference efficiency, and data requirements. It inherits a strong reasoning foundation but uses a hybrid approach to combine the strengths of alternatives while mitigating their drawbacks. Our model defaults to direct inference for perception-focused domains where reasoning adds latency without improving accuracy, avoiding unnecessary verbosity and reducing inference costs, and it invokes longer reasoning paths for domains, such as math and science, that benefit from structured multi-step reasoning.

Implementation details. Our model is trained with SFT, where reasoning samples include `<think>...</think>` sections with chain-of-thought reasoning before the final answer, covering domains like math and science. Non-reasoning samples are tagged to start with a `<nothink>` token, signaling a direct response, and cover perception-focused tasks such as captioning, grounding, OCR, and simple VQA. Reasoning data comprises

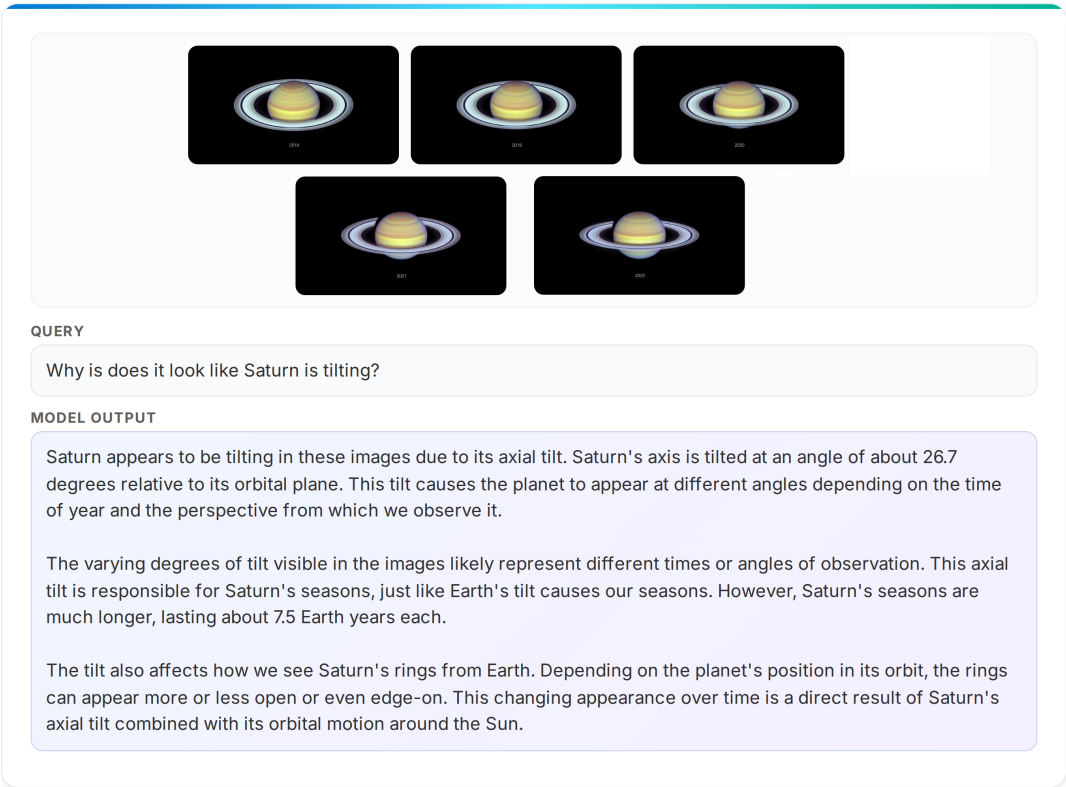


Figure 5: Phi-4-reasoning-vision-15B can interpret sequences of images, here reasoning about the changing appearance of Saturn's rings across multiple frames.

approximately 20% of the total mix. Starting from a reasoning-capable backbone means this data grounds existing reasoning in visual contexts rather than teaching it to reason from scratch.

Limitations and open questions. This approach is not without limitations. The balance between modes is a direct function of design choices we made, informed by recent literature and observed model behavior during training. However, the boundary between modes can be imprecise as it is learned implicitly from the data distribution. Our model allows users to control this behavior through explicit prompting with `<think>` or `<nothink>` tokens when they want to override the default reasoning behavior. The 20/80 reasoning-to-non-reasoning data split may not be optimal for all domains or deployment contexts. Determining the ideal data balance, and ensuring that the model switches appropriately between modes, remains an open research problem.

We view this mixed approach not as a definitive solution, but as one well-motivated point in the design space for balancing latency, accuracy, and flexibility in multimodal systems.

5 Applications

Phi-4-reasoning-vision-15B is a high-performing model across many vision-language tasks. It sees and understands the world by looking at a photo, document, chart, or screen and making sense of it. In practice that covers an enormous range of applications—just a few examples include: describing images and answering questions about them, interpreting changes and trends in image sequences, and recognizing objects, landmarks, and transcribing text. Several examples are shown in Figure 1 and 5.

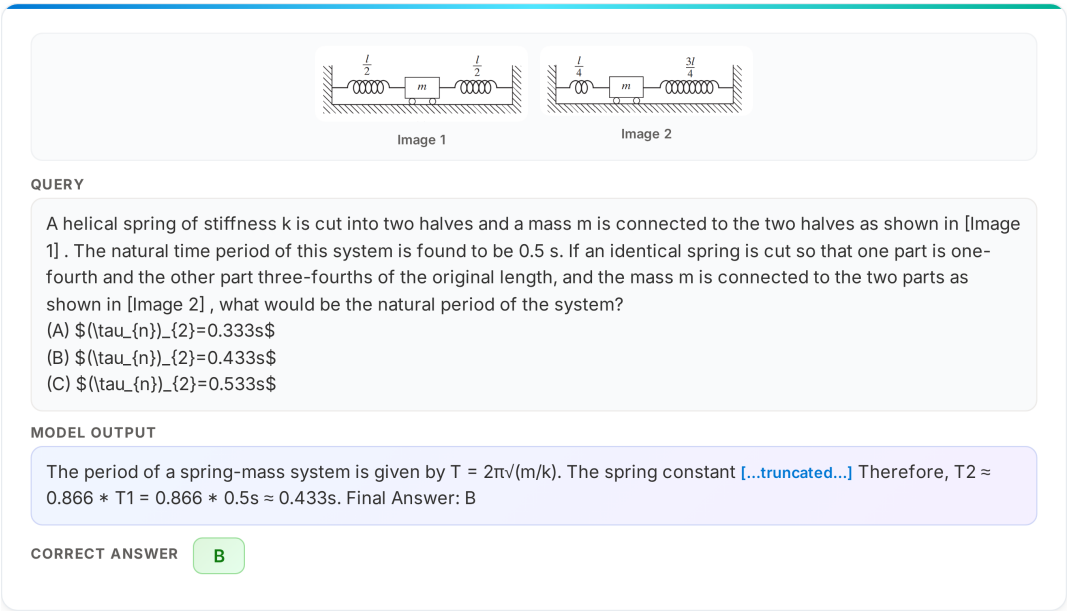


Figure 6: Phi-4-reasoning-vision-15B excels at math and science reasoning, correctly solving a multi-part spring-mass physics problem presented with diagrams.

5.1 Scientific and Mathematical Reasoning

In addition to general vision and language tasks, Phi-4-reasoning-vision-15B was designed to excel at tasks that combine visual input with structured inference, such as solving math problems presented in visual form (e.g., handwritten or diagram-based questions), extracting and reasoning over quantitative information in documents and charts, and supporting multi-step reasoning in educational or scientific analysis contexts. Some examples are shown in Figure 6 and 7.

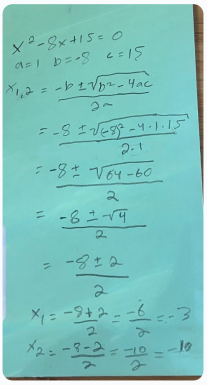
5.2 Computer-Using Agents (CUA)

We trained Phi-4-reasoning-vision-15B to develop capabilities that enable agents to interact with graphical user interfaces (GUIs). The model can interpret screen content and identify appropriate actions. Some examples are shown in Figure 8. With strong high-resolution perception and fine-grained grounding abilities, Phi-4-reasoning-vision-15B provides a strong foundation for building agentic systems. These systems can navigate desktop, web, and mobile environments by detecting and localizing interactive elements such as buttons, menus, and text fields. The model's visual understanding and spatial grounding allow it to reason about interface structure and determine appropriate interactions. Due to its low inference-time needs, it is well-suited for interactive environments where low latency and compact model size are essential.

6 Evaluation

Phi-4-reasoning-vision-15B was evaluated for accuracy and timing using two complementary open-source frameworks to ensure both rigorous and standardized analysis: Eureka ML Insights[1] and VLMEvalKit[2]. We report results on the following benchmarks: AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), HallusionBench (Guan et al., 2024), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), MMStar (Chen et al., 2024), OCRBench (Liu et al., 2024b), and ScreenSpot$_{v2}$sc (Cheng et al., 2024). Accuracy results are presented for our model and compared to several current, open-weight non-thinking and thinking models in Tables 5 and 6, respectively.

[1] https://github.com/microsoft/eureka-ml-insights
[2] https://github.com/open-compass/VLMEvalKit



QUERY

Check my work and see if it is correct. If I'm wrong, answer by pointing out my specific mistake followed by the correct solution.

THINKING

<think>
Let's go through your work step by step and spot the error. 1. You correctly identified the quadratic as $x^2 - 8x + 15 = 0$ with a = 1, b = −8, c = 15. 2. You wrote the quadratic formula as $x = [-b \pm \sqrt{(b^2 - 4ac)}]$ [...truncated...] = [8 ± √4] / 2 = [8 ± 2] / 2 x_1 = (8 + 2) / 2 = 10 / 2 = 5 x_2 = (8 − 2) / 2 = 6 / 2 = 3 So the correct roots are x = 5 and x = 3.
</think>

MODEL OUTPUT

Your work contains one sign error in the substitution step. You wrote $-8 \pm \sqrt{(b^2 - 4ac)}$ instead of $8 \pm \sqrt{(b^2 - 4ac)}$. Because b = −8, −(−8) = +8.

Here is the corrected solution:

1. Start with the quadratic equation:
$x^2 - 8x + 15 = 0$
Identify a = 1, b = −8, c = 15.

2. Apply the quadratic formula:
$x = [-b \pm \sqrt{(b^2 - 4ac)}] / (2a)$

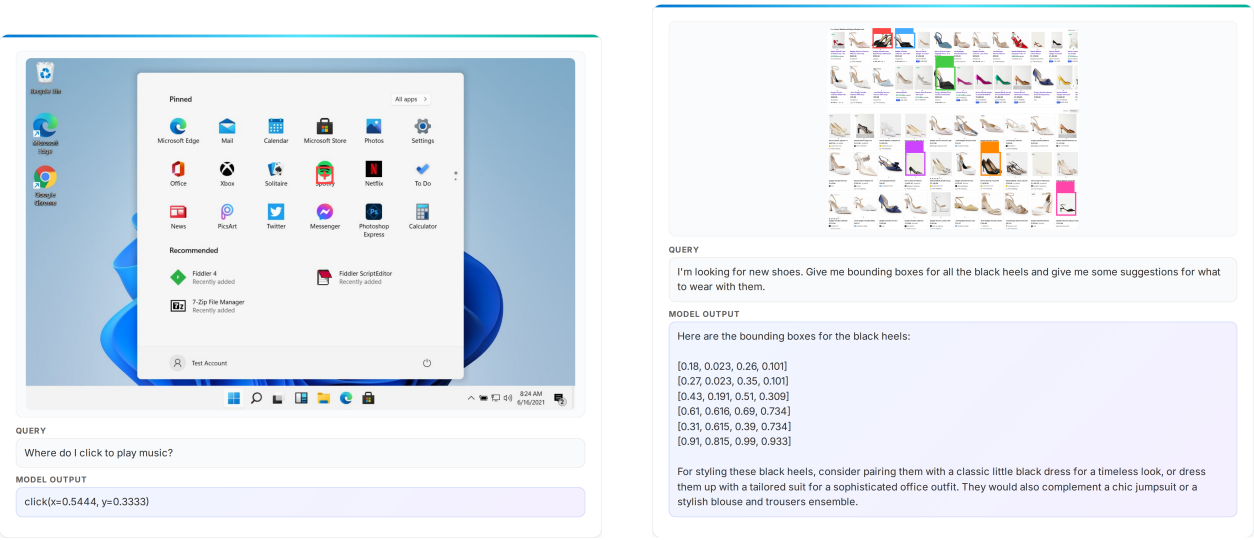
3. Substitute the values:
$x = [-(-8) \pm \sqrt{((-8)^2 - 4 \cdot 1 \cdot 15)}] / (2 \cdot 1)$
$= [8 \pm \sqrt{(64 - 60)}] / 2$
$= [8 \pm \sqrt{4}] / 2$
$= [8 \pm 2] / 2$

4. Split into two solutions:
$x_1 = (8 + 2) / 2 = 10 / 2 = 5$
$x_2 = (8 - 2) / 2 = 6 / 2 = 3$

The correct roots are x = 5 and x = 3.

Figure 7: Phi-4-reasoning-vision-15B can help with written math problems, identifying a sign error in a handwritten quadratic formula solution and providing a corrected step-by-step derivation.

As shown in Tables 5 and 6, our model balances thinking and non-thinking performance—on average showing better accuracy in the default mixed-reasoning behavior than when forcing thinking vs. non-thinking. Only in a few cases does forcing a specific mode improve performance (MathVerse and MMMU_{VAL} for thinking



(a) GUI grounding on a Windows desktop.

(b) Object grounding in a product catalog.

Figure 8: Phi-4-reasoning-vision-15B can help navigate computer UIs, grounding interactive elements on desktop interfaces and localizing objects across dense visual layouts.

and ScreenSpot$_{v2}$ for non-thinking).

Timing experiments. To produce the accuracy-vs-compute plots in Figure 2, we randomly sampled 100 examples from each of four benchmarks—ChartQA$_{TEST}$, MathVista$_{MINI}$, MMMU$_{VAL}$, and ScreenSpot—and measured wall-clock latency and output token counts for every model. All timing experiments were conducted using Eureka ML Insights on NVIDIA H100 GPUs using a single thread with no concurrency and a batch size of one, in order to obtain the most accurate estimate of per-query latency similar to what a user would experience in an interactive setting. We initially tried with vLLM to performing timing experiments using the recommended parameters for each model; while this increased throughput, it also increased per-query latency, giving an inflated estimate of the interactive timing we aimed to measure.

Compared to recent popular, open-weight models, Phi-4-reasoning-vision-15B provides a desirable trade-off between accuracy and cost (as a function of inference time compute and output tokens).

Note: All numbers here are the result of running benchmarks ourselves and may be lower than other previously shared numbers. Instead of quoting leaderboards, we performed our own benchmarking, so we

Table 5: Accuracy comparisons relative to popular open-weight, non-thinking models.

Benchmark	Phi-4-reasoning-vision-15B	Phi-4-reasoning-vision-15B –force nothink	Phi-4-mm-instruct	Kimi-VL-A3B-Instruct	gemma-3-12b-it	Qwen3-VL-8B-Instruct-4K	Qwen3-VL-8B-Instruct-32K	Qwen3-VL-32B-Instruct-4K	Qwen3-VL-32B-Instruct-32K
AI2D$_{TEST}$	84.8	84.7	68.6	84.6	80.4	82.7	83	84.8	85
ChartQA$_{TEST}$	83.3	76.5	23.5	87	39	83.1	83.2	84.3	84
HallusionBench	64.4	63.1	56	65.2	65.3	73.5	74.1	74.4	74.9
MathVerse$_{MINI}$	44.9	43.8	32.4	41.7	29.8	54.5	57.4	64.2	64.2
MathVision$_{MINI}$	36.2	34.2	20	28.3	31.9	45.7	50	54.3	60.5
MathVista$_{MINI}$	75.2	68.7	50.5	67.1	57.4	77.1	76.4	82.5	81.8
MMMU$_{VAL}$	54.3	52	42.3	52	50	60.7	64.6	68.6	70.6
MMStar	64.5	63.9	45.9	60	59.4	68.9	69.9	73.7	74.3
OCRBench	76	75.6	62.6	86.5	75.3	89.2	90	88.5	88.5
ScreenSpot$_{v2}$	88.2	88.3	28.5	89.8	3.5	91.5	91.5	93.7	93.9

Table 6: Accuracy comparisons relative to popular open-weight, thinking models.

Benchmark	Phi-4-reasoning-vision-15B	Phi-4-reasoning-vision-15B –force thinking	Kimi-VL-A3B-Thinking	gemma-3-12b-it	Qwen3-VL-8B-Thinking-4K	Qwen3-VL-8B-Thinking-40K	Qwen3-VL-32B-Thinking-4K	Qwen3-VL-32B-Thinking-40K
$AI2D_{TEST}$	84.8	79.7	81.2	80.4	83.5	83.9	86.9	87.2
$ChartQA_{TEST}$	83.3	82.9	73.3	39	78	78.6	78.5	79.1
HallusionBench	64.4	63.9	70.6	65.3	71.6	73	76.4	76.6
$MathVerse_{MINI}$	44.9	53.1	61	29.8	67.3	73.3	78.3	78.2
$MathVision_{MINI}$	36.2	36.2	50.3	31.9	43.1	50.7	60.9	58.6
$MathVista_{MINI}$	75.2	74.1	78.6	57.4	77.7	79.5	83.9	83.8
$MMMU_{VAL}$	54.3	55	60.2	50	59.3	65.3	72	72.2
MMStar	64.5	63.9	69.6	59.4	69.3	72.3	75.5	75.7
OCRBench	76	73.7	79.9	75.3	81.2	82	83.7	85
$ScreenSpot_{v2}$	88.2	88.1	81.8	3.5	93.3	92.7	83.1	83.1

could understand scaling performance as a function of output token counts for related models. We made our best effort to run fair evaluations and used recommended evaluation platforms with model-specific recommended settings and prompts provided for all third-party models. For Qwen models we use the recommended token counts and also ran evaluations matching our max output token count of 4096. For Phi-4-reasoning-vision-15B, we used our system prompt and chat template but did not do any custom user-prompting or parameter tuning, and we ran all evaluations with temperature $= 0.0$, greedy decoding, and 4096 max output tokens. These numbers are provided for comparison and analysis rather than as leaderboard claims. For maximum transparency and fairness, we will release all our evaluation logs publicly.

7 Safety

As with other Phi models, Phi-4-reasoning-vision-15B was developed with safety as a core consideration throughout training and evaluation. The model was trained on a mixture of public safety datasets and internally generated examples designed to elicit behaviors the model should appropriately refuse, in alignment with Microsoft's Responsible AI Principles. These safety-focused training signals help the model recognize and decline requests that fall outside intended or acceptable use.

Specifically, Phase 3 of training (Section 2.3) incorporates dedicated open-source, responsible AI (RAI) data, including Hateful Memes (Kiela et al., 2021), VLGuard (Zong et al., 2024), Think-in-Safety (Lou et al., 2025), WildGuard (Han et al., 2024). This data covers a range of safety-relevant scenarios such as hateful content detection, refusal of harmful requests, and safe reasoning under adversarial prompts.

Phi-4-Reasoning-Vision-15B's safety was evaluated using both quantitative and qualitative approaches. Automated red teaming was performed on Azure to assess safety risks across multiple risk categories, including disallowed content (sexual, violent, hateful, or self-harm content), copyright content and intellectual property, and jailbreak susceptibility. The evaluation assessed the model's groundedness and its tendency to generate fabricated or misleading information. The safety evaluation built upon the established practices from the Phi-4-Reasoning model's safety assessment. The multimodal nature of the model introduces additional safety considerations around visual content interpretation, and evaluations were conducted to assess the model's behavior when presented with potentially harmful or misleading visual inputs.

Evaluation	Description	Defect Rate
Text to Text Safety	Automated content safety evaluation measuring safety policies	1.4%
Image to Text Safety	Automated content safety evaluation measuring safety policies	4.5%

8 Limitations

While Phi-4-reasoning-vision-15B achieves strong results relative to its size and compute budget, several limitations should be noted:

- Larger proprietary models outperform on broad, unconstrained vision–language benchmarks and generalist multimodal tasks. Phi-4-reasoning-vision-15B is competitive with open-weight models of similar size, and achieves state-of-the-art accuracy relative to training and inference-time compute and tokens—less compute and fewer tokens translates to less waiting and reduced cost.
- The learned switching between reasoning and non-reasoning modes is not always optimal. In some cases, the model may reason when a direct response would suffice, or respond directly when reasoning would be beneficial. Explicit prompting with `<think>` or `<nothink>` tokens can be used to override the default behavior when needed.
- Like many models of its size, Phi-4-reasoning-vision-15B has limitations particularly around extremely detailed or nuanced understanding of images. Users should verify critical outputs, especially for fine-grained visual details.

9 Open Release and Community Engagement

Phi-4-reasoning-vision-15B is available on Microsoft Foundry and HuggingFace with additional examples and details on GitHub. For additional guidance on how to use our model properly and safely, please refer to our Model Card.

In line with our goal of supporting future AI development in the community, Phi-4-reasoning-vision-15B is released under a permissive license with model weights, fine-tuning code, and benchmark logs. We plan to release a portion of our training data in the coming months. We intend this release to complement existing work by providing concrete artifacts that help close gaps in understanding how compact multimodal reasoning models can be built and studied.

10 Looking Forward

Smaller vision–language models with selective, task-aware reasoning offer one promising direction for making multimodal systems more practical and accessible. We present our model and its learnings to inform ongoing research in multimodal modeling, computer-using agents, and mathematical scientific reasoning.

We hope these details are useful to researchers exploring similar tradeoffs and invite critical evaluation, replication, and extension by the community.

Acknowledgements

We thank Rachel Ward for her extensive work on data collection and curation. We thank the GenDatasets, PhiGround, SimCity, and Fara-7B efforts for invaluable training data. We thank Harkirat Behl, Mojan Javaheripi, and Suriya Gunasekar for providing us with Phi-4 checkpoints and guidance on training with Phi models. We additionally thank Sahaj Agarwal, Ahmed Awadallah, Qi Dai, Gustavo de Rosa, Rafah Hosn, Ece Kamar, Piero Kauffmann, Yash Lara, Chong Luo, Caio César Teodoro Mendes, Akshay Nambi, Craig Presti, Matthew Rosoff, Corby Rosset, Marco Rossi, Kashyap Patel, Adil Salim, Sidhartha Sen, Shital Shah, Pratyusha Sharma, Alexey Taymanov, Vibhav Vineet, John Weiss, Spencer Whitehead, the AI Frontiers Team and Leadership, and Microsoft Research Leadership, for their valuable help, insightful discussions, and continued support throughout this work.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025. URL https://arxiv.org/abs/2504.21318.

AI-MO Team. NuminaMath. https://huggingface.co/AI-MO, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Vidhisha Balachandran, Jingya Chen, Neel Joshi, Besmira Nushi, Hamid Palangi, Eduardo Salinas, Vibhav Vineet, James Woffinden-Luey, and Safoora Yousefi. Eureka: Evaluating and understanding large foundation models, 2024. URL https://arxiv.org/abs/2409.10566.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2018.07.011. URL https://www.sciencedirect.com/science/article/pii/S0893608018302107.

Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does throwing away data improve worst-group error? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. URL https://arxiv.org/abs/2311.12793.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents, 2024. URL https://arxiv.org/abs/2401.10935.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi,

Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Eedi. Eedi—mining misconceptions in mathematics. https://eedi.com, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. URL https://arxiv.org/abs/2310.14566.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023. URL https://arxiv.org/abs/2306.11644.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024. URL https://arxiv.org/abs/2406.18495.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective, 2024. URL https://arxiv.org/abs/2402.11530.

HuggingFaceM4 Team. Docmatix: A massive dataset for document visual question answering. *Hugging Face*, 2024a.

HuggingFaceM4 Team. WebSight: A synthetic dataset for improving code generation of screenshot-to-code models. https://huggingface.co/datasets/HuggingFaceM4/WebSight, 2024b.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy, 2022. URL https://arxiv.org/abs/2110.14503.

Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhu Chen. Visualwebinstruct: Scaling up multimodal instruction data through web search, 2025. URL https://arxiv.org/abs/2503.10582.

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. Visonlyqa: Large vision language models still struggle with visual perception of geometric information, 2025. URL https://arxiv.org/abs/2412.00947.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. URL https://arxiv.org/abs/1603.07396.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2021. URL https://arxiv.org/abs/2005.04790.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL http://dx.doi.org/10.1007/s11263-020-01316-z.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, Deli Zhao, Wei Lu, Yu Rong, Aixin Sun, and Shijian Lu. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources, 2025. URL https://arxiv.org/abs/2509.21268.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024b. URL https://arxiv.org/abs/2407.07895.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use, 2025. URL https://arxiv.org/abs/2504.07981.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Junpeng Liu, Tianyue Ou, Yifan Song, Yuxiao Qu, Wai Lam, Chenyan Xiong, Wenhu Chen, Graham Neubig, and Xiang Yue. Harnessing webpage uis for text-rich visual understanding, 2024a. URL https://arxiv.org/abs/2410.13824.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024b. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2025a. URL https://arxiv.org/abs/2412.04468.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4122–4134, June 2025b.

Xinyue Lou, You Li, Jinan Xu, Xiangyu Shi, Chi Chen, and Kaiyu Huang. Think in safety: Unveiling and mitigating safety alignment collapse in multimodal large reasoning model, 2025. URL https://arxiv.org/abs/2505.06538.

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, 2024. URL https://arxiv.org/abs/2310.02255.

U.-V. Marti and H. Bunke. The IAM-database: An English sentence database for off-line handwriting recognition. *International Journal on Document Analysis and Recognition*, 2002.

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, 2022. URL https://arxiv.org/abs/2203.10244.

Juan Rodriguez, Xiangru Jian, Siba Smarak Panigrahi, Tianyu Zhang, Aarash Feizi, Abhay Puri, Akshay Kalkunte, François Savard, Ahmed Masry, Shravan Nayak, Rabiul Awal, Mahsa Massoud, Amirhossein Abaskohi, Zichao Li, Suyuchen Wang, Pierre-André Noël, Mats Leon Richter, Saverio Vadacchino, Shubham Agarwal, Sanket Biswas, Sara Shanian, Ying Zhang, Noah Bolger, Kurt MacDonald, Simon Fauvel, Sathwik Tejaswi, Srinivas Sunkara, Joao Monteiro, Krishnamurthy DJ Dvijotham, Torsten Scholak, Nicolas Chapados, Sepideh Kharagani, Sean Hughes, M. Özsu, Siva Reddy, Marco Pedersoli, Yoshua Bengio, Christopher Pal, Issam Laradji, Spandana Gella, Perouz Taslakian, David Vazquez, and Sai Rajeswar. Bigdocs: An open dataset for training multimodal models on document and code tasks, 2025. URL https://arxiv.org/abs/2412.04626.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025a. URL https://arxiv.org/abs/2503.19786.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran

He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinhao Li, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yuhao Dong, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-vl technical report, 2025b. URL https://arxiv.org/abs/2504.07491.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. URL https://arxiv.org/abs/2402.14804.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms, 2023. URL https://arxiv.org/abs/2312.14135.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction, 2025. URL https://arxiv.org/abs/2412.04454.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, and Christopher Clark. Scaling text-rich image understanding via code-guided synthetic multimodal data generation, 2025. URL https://arxiv.org/abs/2502.14846.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Miaosen Zhang, Ziqiang Xu, Jialiang Zhu, Qi Dai, Kai Qiu, Yifan Yang, Chong Luo, Tianyi Chen, Justin Wagle, Tim Franklin, and Baining Guo. Phi-ground tech report: Advancing perception in gui grounding, 2025. URL https://arxiv.org/abs/2507.23779.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. URL https://arxiv.org/abs/2403.14624.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models, 2024. URL https://arxiv.org/abs/2402.02207.

A Open-Source Training Data

Table 7: Open-Source Training Data Sources for Stages 1–3.

Stage	Category	Datasets
Stage 1: MLP Pretraining	Image–Text Alignment	Bunny (He et al., 2024)
Stage 2: Single-Image Instruction Tuning	Caption	Bunny (He et al., 2024) Recaptioned, Pixmo (Deitke et al., 2024), LLaVA-OneVision (Li et al., 2024a)
	Diagram & Chart QA	LLaVA-OneVision (Li et al., 2024a), CoSyn (Yang et al., 2025)
	Document QA	Docmatix (HuggingFaceM4 Team, 2024a), LLaVA-OneVision (Li et al., 2024a)
	Object Detection	Open Images (Kuznetsova et al., 2020)
	OCR	LLaVA-OneVision (Li et al., 2024a), IAM (Marti and Bunke, 2002)
	Perception	LLaVA-OneVision (Li et al., 2024a), VisOnlyQA (Kamoi et al., 2025)
	Text	NuminaMath (AI-MO Team, 2024), OpenThoughts (Guha et al., 2025)
	VQA	Bunny (He et al., 2024), LLaVA-OneVision (Li et al., 2024a), LLaVA-NeXT (Li et al., 2024a), ShareGPT4V (Chen et al., 2023), Pixmo (Deitke et al., 2024)
	Math: OCR	NuminaMath (AI-MO Team, 2024)
	Math: Problem	LLaVA-OneVision (Li et al., 2024a), NuminaMath (AI-MO Team, 2024), MMR1 (Leng et al., 2025), Eedi (Eedi, 2024)
	CUA: General	AGUVis (Xu et al., 2025), MultiUI (Liu et al., 2024a), Pixmo (Deitke et al., 2024), CoSyn (Yang et al., 2025), BigDocs (Rodriguez et al., 2025)
	CUA: Grounding	PhiGround (Zhang et al., 2025), SeeClick (Cheng et al., 2024)
	CUA: HTML	WebSight (HuggingFaceM4 Team, 2024b)
	RAI	Hateful Memes (Kiela et al., 2021), Think-in-Safety (Lou et al., 2025), WildGuard (Han et al., 2024)
Stage 3: Long Context, Mulit-Image, and RAI	Caption	Bunny (He et al., 2024)
	Document QA	Docmatix (HuggingFaceM4 Team, 2024a)
	VQA	M4-Instruct (Li et al., 2024b)
	Math & Science	VisualWebInstruct (Jia et al., 2025)
	CUA	AGUVis (Xu et al., 2025)
	RAI	Hateful Memes (Kiela et al., 2021), VLGuard (Zong et al., 2024), Think-in-Safety (Lou et al., 2025), WildGuard (Han et al., 2024)