

# MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data and Training Recipes

## Supplementary Material

### A. Implementation Details

During GRPO training, each batch consists of 128 prompts with 8 responses each, and a max response length of 8192 tokens to support detailed reasoning. Rollouts use a temperature of 1.0, with 50% of prompts assigned to long reasoning mode. We remove the thinking process of long reasoning rollouts before pass them to the reward model to overcome the out-of-distribution problem, since the reward model has not been trained on long reasoning samples. We use a fixed learning rate of  $1 \times 10^{-6}$  throughout RL. In the RLAIF-V [96] stage, we use a global batch size of 256, learning rate of  $1 \times 10^{-6}$ , and  $\beta = 0.1$  for 400 steps.

### B. Training Data

#### B.1. Pre-training Data

The summarization of our pre-training data is shown in Table 7.

**Image Caption Data.** We combine large-scale public datasets (LAION-2B [70], COYO [8], etc.) with curated Chinese image-text pairs crawled from the web. We filter out low-resolution images and remove irrelevant image-text pairs with CLIP [67]. To enrich alt-text descriptions, we employ a Capsfusion-based [95] re-captioning process on a subset to generate fluent and factually complete captions. In this way, we formulate the valuable world knowledge in raw captions into more fluent natural language. We employ an MLLM to tag images with concept labels and ensure a balanced distribution across languages and long-tail concepts.

**Image-Text Interleaved Data.** Sourced from Common Crawl [1], OmniCorpus [43] and MINT-1T [2], image-text interleaved data is crucial for in-context learning and multi-image understanding capabilities. We apply filtering to ensure quality, removing samples with broken images or imbalanced image-text ratios. We further use relevance filtering to ensure meaningful multimodal associations, and employ knowledge density filtering to select a high-quality subset for the final decay phase of pre-training.

**OCR Data.** We synthesize OCR data to enhance the basic text recognition capability during the early pre-training stage. We render text on natural scenes with various combinations of color and font following [30], and also render real-world HTML sources into images.

**Document Data.** We collect documents, including scientific papers, academic reports, textbooks, etc., from the web. This data exhibits high knowledge density and contains visually complex layouts.

**Video Caption Data.** We aggregate several public datasets [5, 58, 90], and supplement them with more detailed in-house video captions. This diverse collection supports the development of temporal visual reasoning capabilities essential for video comprehension.

Table 7. Pre-training stages and data composition.

Stage	Sub-stage	New Data Types	Tokens
Cross-Modal Alignment	-	Image Caption	30B
Perceptual Foundation	-	OCR	140B
Holistic Multimodal Tuning	Stable	Document Video Caption Text-only	450B
	Decay	Instruction Interleaved	

Table 8. Supervised fine-tuning data composition.

Stage	Data Types	Tokens
General Supervised Fine-Tuning	General VQA Problem-solving Document Parsing & OCR Multi Image Understanding Video Caption & QA Instruction Following Text-only	80B
Long-CoT & 3D-Resampler Cold-start	General SFT Long-CoT Long Video Fine-Grained Video	10B

#### B.2. Supervised Fine-tuning Data

We summarize our SFT data in Table 8.

**STEM Data.** To enhance STEM reasoning, we curate a dataset of high-school and undergraduate level multidisciplinary problems from the web, covering physics, chemistry, biology, finance, computer science, etc. To ensure the data quality, we implement a two-stage filtering process. First, we only keep samples that exhibit high visual dependency (i.e., not solvable without image information).

Table 9. Data composition for reinforcement learning.

Accuracy Reward Types	Domain	Size
RLVR	Math	158k
	Table	54k
	Document	39k
	Chart	52k
	General Reasoning	79k
RLPR	Table	2k
	Document	3k
	Chart	23k
	General Reasoning	187k
Instruct-Following	Text-only Instructions	60k
<b>Total</b>	-	<b>658k</b>

Second, we perform a consistency check to validate the correctness of answers. For each remaining sample, we collect a clean reasoning process via rejection sampling with a powerful MLLM.

**Long-tail Knowledge Data.** To address the long-tail problem where models often fail on less common topics, we incorporate long-tail knowledge from Wikipedia [86] to synthesize high-quality multimodal instruction-following data. Specifically, for each entity page, we construct multimodal instructions and answers using strong MLLMs and keep samples with high visual dependency.

**Long-CoT Data.** Long-CoT data enables the model to acquire the necessary reasoning patterns for the long reasoning mode. Our data comes from OpenThoughts [28] and an in-house pipeline. We identify challenging prompts by filtering for those on which our early-stage models struggle. Our pilot studies show that focusing on challenging problems is the key to developing robust reasoning capabilities rather than memorizing trivial patterns. Each response then undergoes a multistage validation: we verify its correctness, assess trustworthiness with claim-level factual verification using RLAIIF-V [96], and filter out meaningless repetition. Finally, validated responses are augmented through rewriting to enhance diversity.

#### Takeaway

Filtering out easy prompts and focusing on challenging problems is crucial for effective Long-CoT warm-up.

### B.3. Reinforcement Learning Data

Our RL data contains high-quality samples that span four key domains. Each subset underwent a rigorous, human-in-the-loop cleaning and deduplication process. The summarization of RL data composition is shown in Table 9.

**Mathematics.** We collect multimodal math problems, which require the integration of visual perception and logi-

cal reasoning, from academic sources [20, 25, 44]. We observe that many open-source datasets contain severe label errors and adopt a thorough cleaning process to clean them.

**Documents, Tables and Charts.** To improve reasoning on perceptually complex scenarios, we curate a diverse mix of real-world datasets [12, 31, 50, 63, 107] and synthetic datasets [36, 37, 42] to improve the coverage of domains.

**General Reasoning.** To further improve general reasoning capabilities, we assemble a diverse collection of problems covering logical and multi-disciplinary reasoning tasks from VisualWebInstruct [34] and additional web resources. These data exhibit a more complex reference answer style, and many of the problems have more than one sub-question.

**Instruct Following.** We incorporate text-only instructions from the Llama-Nemotron-Post-Training Dataset [6] and the MulDimIF dataset [93]. We observe that the textual instruction-following improvement generalizes well to multimodal instructions.

## C. Related Works

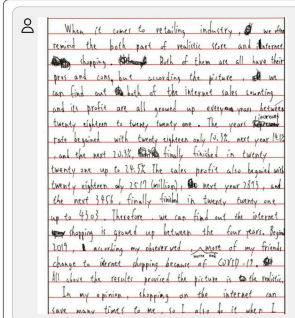
### C.1. Document Data Utilization

Document data is a crucial component for acquiring world knowledge and developing interleaved multimodal understanding for language models [57, 80]. Public PDF documents contain dense academic information and structured knowledge, serving as a rich source for MLLMs to learn generalizable concepts and improve visual alignment [88, 109]. However, converting PDF data into clean and logically coherent token sequences suitable for training is a non-trivial task [40]. Previous MLLM training and data curation practices [26, 40, 49, 89, 109] typically collect public PDF data and rely on standard parsing tools for content extraction. This process usually necessitates extensive and labor-intensive post-processing to filter out noise. Although various PDF parsing pipelines [17, 18, 76, 79, 102, 105] and end-to-end tools [9, 11, 64, 78, 79, 84, 85] exist to assist in this process, driving increasing research into quantifying document understanding abilities [62, 108], significant limitations remain, such as misaligned tables and hallucinated OCR [40, 59]. Accurately converting PDF data into high-quality interleaved training sequences is still challenging, and existing methods often introduce substantial complexity to the data cleaning process. Instead, our proposed unified learning paradigm eliminates the need for explicit external parsers and directly learns from the document image, bypassing the heavy engineering cost and noisy data problem.

### C.2. Video Encoding

The rapid progress of MLLMs has significantly improved video comprehension capabilities [4, 13, 41, 68, 75, 82, 92,



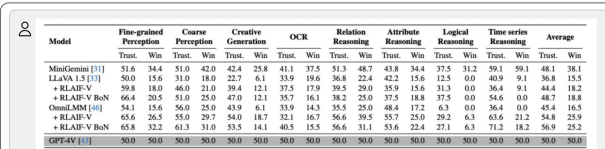


**Extract what's shown in the image, return tables in HTML.**

**MiniCPM-V 4.5**

When it comes to retailing industry, we often remind the both part of realistic store and internet shopping. Both of them are all have their pros and cons, but according to the picture, we can find out both of the internet sales counting and its profit are all grewed up every year between twenty eighteen to twenty twenty one. The years () rate began with twenty eighteen only 10.3%, next year 14.1%, and the next 20.3%, finally finished in twenty twenty one up to 24.5%. The sales profit also began with twenty eighteen only 2517 (million), next year 2893, and the next 3456, finally finished in twenty twenty one up to 4303. Therefore, we can find out the internet shopping is grewed up between the four years. Begun 2019, according my observed, () more of my friends change to internet shopping because of COVID-19. All above the results provided the picture is the realistic. In my opinion, shopping on the internet can save many times to me, so I also do it when I

Figure 8. A case of accurate and robust recognition of handwritten text containing scribbles.



**Extract what's shown in the image, return tables in HTML.**

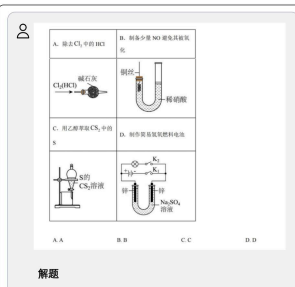
**MiniCPM-V 4.5**

Model	Fine-grained Perception		Course Perception		Creative Generation		OCR		Relation Reasoning		Attribute Reasoning		Logical Reasoning		Time series Reasoning		Average	
	Trust.	Win.	Trust.	Win.	Trust.	Win.	Trust.	Win.	Trust.	Win.	Trust.	Win.	Trust.	Win.	Trust.	Win.	Trust.	Win.
MiniGemini [11]	51.6	34.4	51.0	42.0	42.4	25.8	41.1	37.5	51.3	48.7	43.8	34.4	37.5	31.2	59.1	48.1	48.1	38.1
LLaVA-1.5 [13]	50.0	15.6	31.0	18.0	22.7	6.1	33.9	19.6	36.8	22.4	42.2	15.6	12.5	0.0	40.9	9.1	36.8	15.5
+RLAIF-V	59.8	18.0	46.0	21.0	39.4	12.1	37.5	17.9	39.5	29.0	35.9	15.6	31.3	0.0	36.4	9.1	44.4	18.2
+RLAIF-V BoN	66.4	20.5	51.0	25.0	47.0	12.1	35.7	16.1	38.2	25.0	37.5	18.8	37.5	0.0	54.6	0.0	48.7	18.8
OsmoLLM [6]	54.1	15.6	56.0	25.0	43.9	6.1	33.9	14.3	35.5	25.0	48.4	17.2	6.3	0.0	36.4	0.0	45.4	16.5
+RLAIF-V	65.6	26.5	55.0	29.7	54.0	18.7	32.1	16.7	56.6	39.5	55.7	25.0	29.2	6.3	63.6	21.2	54.8	25.9
+RLAIF-V BoN	65.8	32.2	61.3	31.0	53.5	14.1	40.5	15.5	56.6	31.1	53.6	22.4	27.1	6.3	71.2	18.2	56.9	25.2
GPT-4V [13]	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

Table 7. The trustworthiness win rate / overall win rate of different MLLMs on eight capabilities of ReFoMB test split. Trust.: trustworthiness win rate. Win.: overall win-rate.

Figure 9. A case of table content extraction, where the structure is faithfully recovered.

### D.4. Problem Solving



**MiniCPM-V 4.5**

<think>  
好的，我需要分析每个选项并判断哪个装置是正确的。让我先逐个看每个选项。  
.....  
</think>  
为了解决这个问题，我们逐项分析每个装置。  
选项A：要除去Cl<sub>2</sub>中的HCl，应使用饱和食盐水(NaCl)溶液，因为HCl易溶于水，而Cl<sub>2</sub>在饱和食盐水中溶解度很小。碱石灰会与Cl<sub>2</sub>反应，因此A错误。  
选项B：铜与稀硝酸反应生成NO，由于NO不溶于稀硝酸，此装置可防止NO被氧化，因此B正确。  
选项C：乙醇和CS<sub>2</sub>互溶，不能作为萃取剂分离S，正确的萃取方法应使用与CS<sub>2</sub>不互溶的溶剂(如苯或四氯化碳)，因此C错误。  
选项D：氨氧燃料电池中，正极(氧气电极)应在酸性环境中工作，Na<sub>2</sub>SO<sub>4</sub>溶液是中性的，氧气在正极生成OH<sup>-</sup>，而锌与OH<sup>-</sup>反应生成Zn(OH)<sub>2</sub>，导致电池失效，应使用酸性电解质(如稀硫酸)，因此D错误。  
综上所述，正确答案是B。

Figure 10. A case of K12 level chemistry problem solving in Chinese, equipped with the long reasoning mode for comprehensive analysis.

**Answer these questions.**

For Questions 1-5, please use Figure 1 to compute the correct numerical value and enter the value for each question. An ANOVA test output is shown below (Figure 1).

Figure 1. One-way ANOVA Table

One-way ANOVA					
Source	DF	SS	MS	F	P
Factor	3	36.15	?	?	?
Error	?	?	?		
Total	19	198.04			

Degrees of freedom (DF) associated with Error (residual) = ?

**Question Completion Status:**

**QUESTION 2**

Sum of squares (SS) associated with Error (residual) = ?

.....

**QUESTION 3**

Mean squares (MS) associated with Factor = ?

.....

**QUESTION 4**

Mean squares (MS) associated with Error (residual) = ?

.....

**QUESTION 5**

Test statistic (F) = ?

**MiniCPM-V 4.5**

<think>  
Okay, so I need to solve this one-way ANOVA problem. Hmm, let me start by recalling the .....  
</think>

Answers:  
QUESTION 2: 159.89  
QUESTION 3: 12.05  
QUESTION 4: 9.9931 (or ~9.99)  
QUESTION 5: 1.206  
Note: For Question 1 (not explicitly listed above but implied in the context), the degree of freedom for Error is 16.

Figure 11. A case of multi-image undergraduate-level statistical problem solving, where multiple questions are solved in a coherent response.