A Hierarchical Filtering Framework for Curating High-Quality Visual Instruction Data

Yun Zhu[1], **Honglin Lin**[1], **Yu Li**[1], **Chonghan Qin**[1], **Zheng Liu**[1], **Xiaoyang Wang**[1], **Lijun Wu**[1]*
[1]Shanghai Artificial Intelligence Laboratory, {zhuyun,wulijun}@pjlab.org.cn

Abstract

This technical report introduces a hierarchical framework for curating a high-quality subset of 10,000 visual instruction samples from large-scale, multi-source multimodal datasets. The proposed data selection pipeline employs a sequence of filtering stages—fuzzy deduplication, question-length pruning, visual-dependency assessment, and fail-score evaluation—to ensure maximal diversity, informativeness, and reliability. Empirical results demonstrate that models trained on the curated subset achieve substantial gains on mathematics and science visual reasoning benchmarks, validating the effectiveness of our approach in enhancing multimodal instruction tuning. The three curated datasets—comprising 2.5K, 3K, and 10K samples—are publicly available at https://huggingface.co/datasets/Dream000/datasets.

1 Introduction

Large-scale multimodal datasets HuggingFaceM4 (2025) are standard for visual instruction tuning, yet not all samples equally benefit model learning. These datasets are often noisy, redundant, and heterogeneous, with instructions varying in quality, style, and visual relevance. Existing selection methods Yu et al. (2024); Liu et al. (2024), relying on single criteria, struggle to handle multi-source heterogeneity, subtle redundancies, and sample reliability, limiting their ability to curate high-quality subsets.

To address this, we propose a hierarchical selection framework that systematically identifies the most informative and reliable samples from massive multi-source datasets. The framework combines multiple filtering stages—fuzzy deduplication, question-length filtering, visual-dependency assessment, and fail-score evaluation—to ensure selected samples are diverse, informative, and visually grounded.

Applying this framework, we curate three high-quality visual instruction subsets. Experiments show that models trained on these curated datasets consistently outperform those trained on larger, unfiltered corpora across mathematics and science visual reasoning benchmarks. These results demonstrate that hierarchical, multi-criteria selection enhances both efficiency and model performance, providing a practical solution for large-scale multimodal instruction tuning.

2 Data Sources

Four major multimodal datasets were considered:

- **LLaVA-CoT-100K** (Xu et al., 2024): Integrating samples from various visual question answering sources and providing structured reasoning annotations.
- **Multimodal Open R1** (EvolvingLMMs-Lab, 2025): Open-sourced the first batch of 8k multimodal RL training examples focused on Math reasoning.

*Corresponding author.

NeurIPS DCVLR Workshop 2025.

- **FineVision** (HuggingFaceM4, 2025): Collecting over 200 datasets containing 17M images, 89M question-answer turns, and 10B answer tokens, totaling 5TB of high-quality data.
- **WeMath 2.0** (Qiao et al., 2025): Ensuring comprehensive conceptual coverage and enhanced flexibility through a dual-expansion strategy in the mathematics domain.

Specifically, due to the large volume of samples in FineVision, we primarily focus on data from the mathematics and science domains. All data samples were standardized into a unified triplet format `{image, problem, solution}`, represented as $((o_i, q_i, a_i) \in \mathcal{D})$.

3 Hierarchical Data Curation Methodology

Our selection pipeline consists of five hierarchical stages designed to progressively refine the dataset.

3.1 Stage 1: Fuzzy Deduplication

We adopt a fuzzy deduplication strategy based on *MinHash* and *Locality-Sensitive Hashing* (LSH) to eliminate near-duplicate questions. Each question q_i is normalized to lowercase and decomposed into overlapping character-level n-grams (default $n = 3$), forming a set of shingles:

$$S_q = q_i^{(n)} \mid i = 1, \ldots, |q| - n + 1. \tag{1}$$

A MinHash signature $m_q \in \mathbb{R}^N$ with $N = 64$ permutations is generated to approximate the Jaccard similarity between two questions q_i and q_j:

$$\text{Jaccard}(S_{q_i}, S_{q_j}) = \frac{|S_{q_i} \cap S_{q_j}|}{|S_{q_i} \cup S_{q_j}|} \approx \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}[m_{q_i}^{(k)} = m_{q_j}^{(k)}]. \tag{2}$$

All MinHash signatures are inserted into an LSH index with similarity threshold $\tau = 0.85$. For each new question, if $\text{Jaccard}(S_{q_i}, S_{q_j}) < \tau$, it is considered unique and retained; otherwise, it is removed as a near-duplicate. This method achieves scalable and reliable fuzzy deduplication in large multimodal datasets. This process eliminates paraphrased duplicates and near-identical questions, ensuring unique instruction coverage. Deduplication reduced the initial corpus size by approximately 25%.

3.2 Stage 2: Short Question Filtering

Samples containing excessively short or underspecified questions were excluded. Specifically, instructions comprising fewer than eight tokens were discarded to ensure that the remaining data provide sufficient semantic context for meaningful reasoning. This filtering step effectively removes trivial or low-information queries, such as generic visual descriptions and questions like "What is this?" or "How many objects are present?", which are unlikely to require nuanced understanding or reasoning. Approximately 20% of the remaining samples were removed at this stage.

3.3 Stage 3: Visual Dependency Filtering

After removing short or underspecified questions, we further filtered samples to ensure they require genuine visual reasoning. Using `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025), we excluded questions answerable without the corresponding image. Formally, the filtered dataset is defined as:

$$\mathcal{D}_{\text{filtered}} = \{(o_i, q_i, a_i) \in \mathcal{D} \mid \text{Acc}_{\text{Qwen2.5-VL-7B}}(o_1, q_i) > \text{Acc}_{\text{Qwen2.5-VL-7B}}(\varnothing, q_i)\}.$$

Here, $\text{Acc}_{\text{Qwen2.5-VL-7B}}(o_i, q_i)$ denotes the accuracy of the model when provided both the image o_i and question q_i, whereas $\text{Acc}_{\text{Qwen2.5-VL-7B}}(\varnothing, q_i)$ represents the accuracy when the model is given only the question. Only samples for which the presence of the image improves model performance were retained, thereby ensuring that the remaining dataset emphasizes tasks that genuinely depend on visual information. This filtering stage removed approximately 22% of the remaining samples, further enhancing the quality and relevance of the dataset for subsequent visual reasoning tasks.

3.4 Stage 4: Fail Score Computation

Each remaining sample was subjected to five independent rollouts using `Qwen2.5-VL-72B-Instruct` as the generation model. For each rollout, the model generated a predicted response given the instruction and image. The *fail score* F_s is defined as:

$$F_s = 1 - \frac{N_{\text{success}}}{5},$$

where N_{success} denotes the number of generated responses semantically equivalent to the reference ground truth.

The fail score ranges from 0 (perfect consistency) to 1 (complete failure), with intermediate values indicating challenging yet solvable tasks that enhance model robustness. This stage removed approximately 75% of the remaining samples, refining the dataset to focus on tasks that effectively evaluate multimodal reasoning.

3.5 Stage 5: Fail Score Filtering and Ranking

The remaining samples, denoted by the set S, were ranked in descending order according to their fail score F_s. The top $k = 10{,}000$ samples were then selected as the final dataset:

$$\text{Final Selection} = \text{TopK}(S, k, \text{key} = F_s), \tag{3}$$

where $\text{TopK}(\cdot)$ returns the k elements of a set with the highest values according to the specified key function. Here, F_s represents the fail score of each sample, which quantifies the model's inconsistency in generating correct responses across multiple rollouts. This approach ensures that the final dataset emphasizes moderately difficult samples that effectively test and enhance the model's reasoning and visual perception capabilities.

4 Conversation Generation Approach

For selected samples that lack chain-of-thought (CoT) reasoning or contain only brief CoT, we employ the following methods to generate high-quality solutions.

Solution Generation: Based on a set of carefully curated high-quality problems, we employ `qwen3-vl-235b-a22b-thinking` (Team, 2025) for solution distillation. We use the prompt following:

Solution Generation Prompt

You are an expert in visual reasoning (algebra, geometry, number theory, combinatorics, calculus, probability, competition problems).
Your response will be used as a high-quality example to train a new AI model.
Solve the problem efficiently and clearly, extracting information from multimodal inputs (text, photos, charts, graphs, diagrams, tables, handwritten notes, etc.) and producing rigorous, step-by-step solutions.
Remember: information from images is as important as text, and must be integrated.

Strict formatting requirements:
1) Put the final result ONLY inside <answer></answer>.
2) If units are required, include them inside the <answer></answer>.
3) For multiple-choice questions, put only the chosen letter (and nothing else) inside the <answer></answer>.
4) The last line of your response must be exactly: "Therefore, the final answer is <answer>ANSWER</answer>."

Problem: {problem}

Please think step by step, carefully examine the provided image, and provide a clear solution.

The hyperparameters are set to a sampling temperature of 0.6, `top_p` of 0.95, and a maximum generation length of 16,384 tokens.

Output Formatting: The Qwen API returns both the reasoning process and the final answer. We convert the output into the following standardized format:

```
<think>Reasoning Content</think>
... Therefore, the final answer is <answer>ANSWER</answer>.
```

Answer Consistency Filtering: We verify the model's generated answer against the ground truth using an LLM-based consistency check. Only samples where the model's final answer matches the ground truth are retained, ensuring alignment between the reasoning process and the correct answer.

Sample Selection Strategy: Among the correctly answered samples, we prioritize those with longer reasoning chains for training. This helps improve coverage of complex reasoning patterns and enhances the model's ability to handle challenging problems.

5 Computational Resources

The entire pipeline was executed on a distributed GPU cluster with the following configuration:

- **GPUs:** 8 $\times$ NVIDIA A100 (80GB)
- **CPUs:** 2 $\times$ AMD EPYC 7763 (64 cores)
- **Memory:** 1 TB RAM

6 Conclusion

This report introduces a hierarchical data selection framework for constructing a high-quality subset of 10,000 visual instruction samples via a multi-stage filtering pipeline augmented with fail score ranking. The proposed approach systematically eliminates redundancy, low-quality entries, and visually irrelevant samples, while emphasizing moderately challenging tasks that foster comprehensive and robust reasoning capabilities.

References

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

EvolvingLMMs-Lab. 2025. open-r1-multimodal: A fork to add multimodal model training to open-r1. `https://github.com/EvolvingLMMs-Lab/open-r1-multimodal`. `https://github.com/EvolvingLMMs-Lab/open-r1-multimodal` Accessed: 2025-10-09.

HuggingFaceM4. 2025. FineVision: Open Data is All You Need. `https://huggingface.co/spaces/HuggingFaceM4/FineVision`. `https://huggingface.co/spaces/HuggingFaceM4/FineVision` Accessed: 2025-10-09.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. 2024. Less is more: High-value data selection for visual instruction tuning. *arXiv preprint arXiv:2403.09559* (2024).

Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, et al. 2025. We-Math 2.0: A Versatile MathBook System for Incentivizing Visual Mathematical Reasoning. *arXiv preprint arXiv:2508.10433* (2025).

Qwen Team. 2025. Qwen3-VL: The Most Powerful Vision-Language Model in the Qwen Series. `https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef&from=research.latest-advancements-list` Accessed: 2025-10-09.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440* (2024).

Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Bosheng Qin, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. 2024. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness. *arXiv preprint arXiv:2412.06293* (2024).