

FineVision: Open Data Is All You Need

Luis Wiedmann^{TUM} 🧑‍💻 ★ Orr Zohar^{Stanford} 🧑‍💻 ★ Amir Mahla 🧑‍💻 Xiaohan Wang^{Stanford} 🧑‍💻 Rui Li^{Stanford} 🧑‍💻 Thibaud Frere 🧑‍💻
Leandro von Werra 🧑‍💻 Aritra Roy Gosthipaty 🧑‍💻 Andrés Marafioti 🧑‍💻

🧑‍💻 Hugging Face, ^{TUM} Technical University Munich, ^{Stanford} Stanford University
★ Equal Contribution

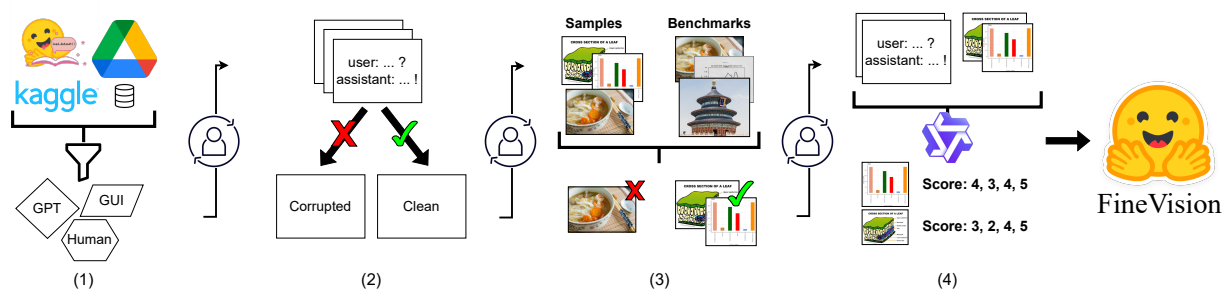


Figure 1 | Pipeline overview. Left to right: (1) ingestion of raw sources; (2) canonicalization and image & text cleaning; (3) de-duplication and test-set decontamination using SSCD embeddings (Pizzi et al., 2022); (4) per-turn quality assessment with LLM/VLM-as-a-judge (Zheng et al., 2023; Wang et al., 2023c). Each stage includes human checkpoints (mapping review, script sign-off, and post-conversion audits) to ensure faithful annotation consumption, consistent quality, and safety.

Abstract



The advancement of vision-language models (VLMs) is hampered by a fragmented landscape of inconsistent and contaminated public datasets. We introduce *FineVision*, a meticulously collected, curated, and unified corpus of 24 million samples—the largest open resource of its kind. We unify more than 200 sources into 185 subsets via a semi-automated, human-in-the-loop pipeline: automation performs bulk ingestion and schema mapping, while reviewers audit mappings and spot-check outputs to verify faithful consumption of annotations, appropriate formatting and diversity, and safety; issues trigger targeted fixes and re-runs. The workflow further applies rigorous de-duplication within and across sources and decontamination against 66 public benchmarks. FineVision also encompasses agentic/GUI tasks with a unified action space; reviewers validate schemas and inspect a sample of trajectories to confirm executable fidelity. Models trained on FineVision consistently outperform those trained on existing open mixtures across a broad evaluation suite, underscoring the benefits of scale, data hygiene, and balanced automation with human oversight. We release the corpus and curation tools to accelerate data-centric VLM research.

Dataset [HuggingFaceM4/FineVision](#)

Blog Post [HuggingFaceM4/FineVision](#)

1 Introduction

The remarkable progress of vision-language models (VLMs) has been fueled by scaling both the capacity of the model and the training data. However, the open research community faces a critical bottleneck: multimodal public data sets are fragmented, inconsistent, and often contaminated. While proprietary models are trained on massive, curated corpora, open alternatives must stitch together many smaller, specialized datasets. This misalignment not only hinders reproducibility, but also widens the performance gap between closed-source and open-source VLMs, limiting the community’s ability to conduct robust, data-centric research.

Historically, early open aggregation efforts began with works like The Cauldron, followed by Cambrian-1 and LLaVA-OneVision (Laurençon et al., 2024; Tong et al., 2024; Li et al., 2024). These efforts were competitive at the time of release and laid crucial groundwork by unifying disparate sources. Subsequently, leading open-source models have shifted toward much larger training mixtures that span hundreds of datasets and often combine open and closed sources; for example, Eagle2 and PerceptionLM (Li et al., 2025b; Cho et al., 2025) each report using on the order of 200 datasets. As the field expands into agentic and GUI-grounded tasks, the need has moved from aggregation to principled, scalable curation. The next frontier of VLM development requires datasets that are not only large-scale but also diverse and are engineered for emerging tasks.

To address this challenge, we introduce **FineVision**, a meticulously engineered corpus of over **24 million samples** with **17 million images**, **89 million turns** and **9.5 billion answer tokens**. Our primary contribution is the collection, rigorous curation, and open release of this dataset, providing a reliable, ready-to-use foundation for training and evaluating VLMs. To enable this, we developed a semi-automated, human-in-the-loop curation workflow that unifies over 200 sources and enforces a consistent chat schema. Automation performs bulk ingestion and schema mapping; reviewers then verify key steps through targeted audits and spot-checks. The pipeline conducts comprehensive cleaning - removing corrupted images and malformed text, validating image-text alignment, and sanitizing unsafe content - alongside rigorous de-duplication within and across sources and decontamination against 66 evaluation benchmarks to protect test integrity. Reviewers audited random samples to confirm faithful consumption of source annotations as well as appropriate formatting and diversity, and requested targeted fixes or re-runs when issues arose; for agentic/GUI data, they validated the unified action schema and inspected a small sample of trajectories to confirm executable fidelity. This review loop was repeated, as necessary, until quality criteria were met.

We validate FineVision through extensive experiments. Models trained on our corpus achieve state-of-the-art results among open-data VLMs, showing significant relative improvements over baselines: **40.7%** over The Cauldron, **12.1%** over Cambrian-1, and **46.3%** over LLaVA-OneVision on an average of 11 benchmarks. These results underscore the value of our principled approach to data hygiene and thoughtful integration.

We release FineVision and its associated resources to the public, aiming to democratize access to high-quality training data and catalyze the next wave of innovation in open VLM development.

2 FineVision Curation

FINEVISION was created through a large-scale data curation effort to address the critical need for diverse, high-quality training data in the open-source VLM community. Our primary contribution is the collection, rigorous curation, and open release of the dataset itself. We unify over 200 public datasets through a semi-automated, human-in-the-loop process into a final corpus of 185 subsets. Automation performs bulk ingestion and schema mapping; reviewers then verify key steps for each dataset. Each source underwent a focused manual audit to accommodate its specific format and annotation style—for example, image QA, multi-image conversations, localized captions, and relational graphs.

We convert each dataset into a standardized chat format suitable for instruction tuning, using multiple conversational templates to ensure stylistic diversity and constructing multi-turn interactions where appropriate. Large language models were used to scale parts of the conversion; however, a reviewer remained in the loop to audit samples, confirm that source annotations were faithfully consumed, and request targeted fixes or re-runs when needed. This review loop was repeated, as necessary, until quality criteria were met. This section details the curation workflow, from data sourcing (Sec. 2.1) and schema unification (Sec. 2.2) to cleaning and

decontamination (Sec. 2.3 and 2.4), and the design choices that enabled this large-scale dataset.

2.1 Data Sources

Our data collection process aimed to be comprehensive, aggregating datasets from wherever they were publicly released by their original authors. We gathered over 200 datasets, sourcing data from a variety of locations:

- **Public Dataset Hubs:** Established platforms like Hugging Face Datasets, which host versioned and documented corpora.
- **Institutional and Cloud Storage:** Publicly shared links on institutional or personal cloud storage (e.g., Google Drive), a common hosting choice for academic releases.
- **Code Repositories:** GitHub repositories where datasets are shared alongside research code, often requiring custom extraction scripts.
- **Direct Web Downloads:** Project websites and other direct download links.

The full per-source breakdown is detailed in the Appendix (Section A.7). After filtering and deduplication, we ended up with 185 subsets.

2.2 From Heterogeneous Annotations to Unified Conversations

Converting over 200 public datasets into a unified format suitable for instruction tuning was a significant engineering challenge. Each dataset arrived with its own annotation schema, task formulation, and data organization - from simple image-caption pairs to complex multi-page document QA with derivations and spatial grounding. This subsection details our systematic approach to transforming this heterogeneous collection into high-quality conversational training data.

Semi-automated conversion pipeline. We developed a hybrid approach combining LLM assistance with human expertise. Using Claude as an agent, we broke down each dataset conversion into manageable subtasks: (1) deep annotation analysis to understand the structure and semantics of each dataset, (2) strategy design to map source annotations to conversational format, (3) script implementation with extensive validation, and (4) quality verification through sampling and then manual human inspection. This approach allowed us to maintain consistency across conversions while adapting to the unique requirements of each dataset. Every conversion script was manually reviewed and tested before full-scale processing.

Human-in-the-loop quality control. We prioritized automation while preserving accountability through targeted oversight. For each dataset, a reviewer (i) assessed the mapping plan and template choices, (ii) examined a dry-run of the converter, and (iii) audited a random sample of outputs to verify complete annotation consumption and appropriate formatting and diversity. When issues arose (e.g., missed fields or brittle templates), reviewers issued focused guidance and re-ran the affected stage. For agentic/GUI data, reviewers additionally validated the unified action schema and inspected a small sample of trajectories to confirm executable fidelity.

Unified conversational schema. All datasets converge to a standardized sample-level representation:

$$\text{sample} = \{\text{images}, \text{texts}, \text{source}, \text{metadata}\}.$$

where `texts` contains a list of conversational turns alternating between `user` and `assistant` roles. Non-conversational sources (e.g., classification datasets) are transformed into natural QA pairs using carefully designed templates. We also experiment with converting single-turn QA datasets into multi-turn conversations by grouping multiple questions about the same image together, to create richer training signals that better leverage each image. We preserve task-specific information in `metadata` for downstream filtering and analysis, including quality ratings, original sources and, confidence scores where available.

Task-specific conversion strategies. We developed six core strategies to handle the diversity of supervision types while preserving their semantic richness. To ensure stylistic diversity, we randomized question templates and answer formats across conversions:

- **Visual QA:** Questions about the same image are grouped into multi-turn conversations. Multiple-choice questions include options in the prompt with answers providing both the selection and rationale. Question phrasings are varied ("What is...", "Can you identify...", "Tell me about...") to avoid templatic patterns.
- **Captioning & Description:** Ground-truth captions are wrapped with randomized instructional prompts ("Describe this image," "What's shown here?", "Provide a detailed description of...") to create natural QA pairs without altering the original descriptions.
- **Grounding & Spatial Relations:** Spatial annotations (e.g., "cat left of dog") become yes/no questions with varied phrasings and explanatory answers. Bounding box coordinates are converted into natural language descriptions of spatial relationships (e.g., left, right, above), while the raw coordinates are normalized to a (cx, cy, w, h) format and preserved in metadata.
- **Document Understanding:** Multi-page documents are processed as image lists with questions threaded into conversations. Answers are enriched with available annotations like derivation steps, supporting facts, and answer types (arithmetic, extractive, etc.).
- **OCR & Transcription:** We generate both exact transcription turns and optional "understanding" turns that explain the content's structure or meaning, particularly useful for mathematical expressions and handwritten text.
- **Classification & Detection:** Binary or categorical labels are converted to decision questions with explanatory answers when auxiliary descriptions are available, maintaining the educational value of the original annotations.

Action-space unification for GUI data. To enable novel capabilities in agentic vision tasks, we include multiple GUI automation datasets where a major challenge is the lack of standardization in action spaces: different sources define heterogeneous function signatures, parameter naming conventions, and action taxonomies. To address this, we built a data transformation pipeline on top of the open-source datasets used in Xu et al. (2025b). Our pipeline includes (i) a parser that extracts and normalizes arbitrary function signatures, ensuring consistent parameter ordering and reconstruction, and (ii) an action conversion module that maps all action representations into a unified schema. This process enforces consistent function and parameter naming, and produces a coherent, typed action schema. Screen coordinates are expressed in normalized form [0,1] to ensure resolution-agnostic training. By unifying the action space, we enable cross-domain training and allow models to learn coherent action patterns across heterogeneous GUI environments (desktop, mobile, or browser). See Appendix (Section A.3) for further details.

2.3 Cleaning

Our workflow includes several automated cleaning and validation steps to handle edge cases common in large-scale data aggregation, such as corrupted files, malformed annotations, and inconsistent text formatting.

Images. We perform automatic image validation, decoding with robust backends to discard undecodable, corrupted, or zero-byte images. We also orient images via EXIF metadata and convert all formats to RGB. Any samples with failed image I/O are dropped from the dataset and we cap the image size at 2048px for the longest side while preserving the aspect ratio.

Text. Text content is normalized to enforce UTF-8 encoding, strip control characters, standardize punctuation and quotes, and remove artifacts like base64 blobs. We collapse repeated tokens (e.g., !!!! → !) and remove turns with empty or degenerate answers (e.g., single-character repeats). To filter outliers and ensure training stability, every turn is capped at a combined question and answer length of 8192 tokens.

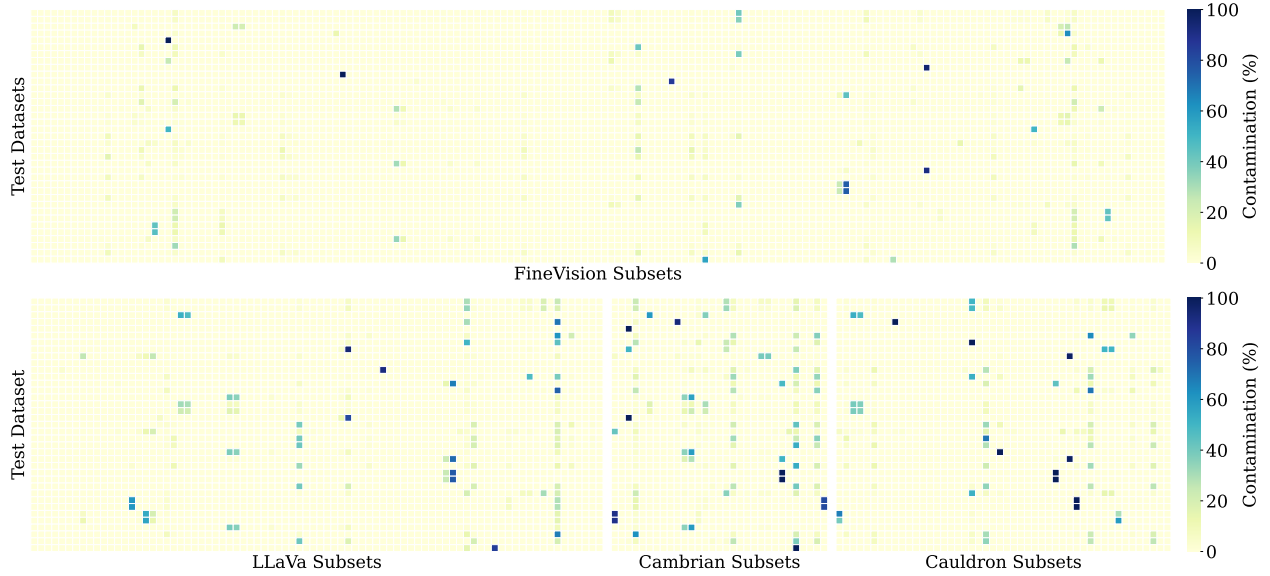


Figure 2 | Decontamination report. Per-benchmark contamination heatmap for FineVision and comparable open-source alternatives (rows: benchmarks, columns: datasets subsets). FineVision’s contamination is sparse and concentrated in a few subsets and benchmarks, and consistently lower than the baselines.

2.4 Near-Duplicate and Contamination Control

We perform hygiene in two stages using self-supervised copy-detection descriptors (SSCD) (Pizzi et al., 2022) and cosine similarity:

1. **Intra-Dataset:** cluster visually near-identical images within FineVision, to merge samples from the same image into a multi-turn conversation or a merged subset.
2. **Test-Set Decontamination:** identifying training images similar to evaluation images from 66 public VLM benchmarks (via embeddings computed once from the same SSCD model), mitigating train–test leakage (Razeghi et al., 2022), see Fig. 2.

All stages share a threshold $\tau = 0.95$ on cosine similarity, erring on the conservative side to reduce false negatives (examples of the different scenarios are in the Appendix, Fig. 8).

Intra-Dataset Duplicates We flag subsets for potential overlap with each other using the SSCD+cosine pipeline and manually inspect them before potentially merging into a single subset (e.g., we merged three commonly found online variants of `ai2d` into a single `ai2d_merged` subset). We additionally experiment with generally merging multiple individual questions for the same image into a multi-turn conversation, but this did not result in improved performance during our ablations.

Contamination Measurement Against Public Benchmarks. Following the same SSCD+cosine protocol, we embed all images from 66 test sets included in `lmms-eval` (EvolvingLMMS-Lab, 2024) and compute their max similarity to each training image. Images with similarity $\geq \tau$ are flagged, and we study the impact of removing them from training, but since this is not a definitive indicator of a contaminated *sample*, we release FineVision in its original form. Detailed description and statistics of the contamination and performance drop across datasets are in the Appendix (Table 4) and the contamination is visualized in Fig. 2. We release both the de-duplication pipeline¹ and the precomputed SSCD embeddings for the used public benchmarks².

¹<https://github.com/huggingface/large-scale-image-deduplication>

²<https://huggingface.co/datasets/HuggingFaceM4/lmms-eval-embeddings>

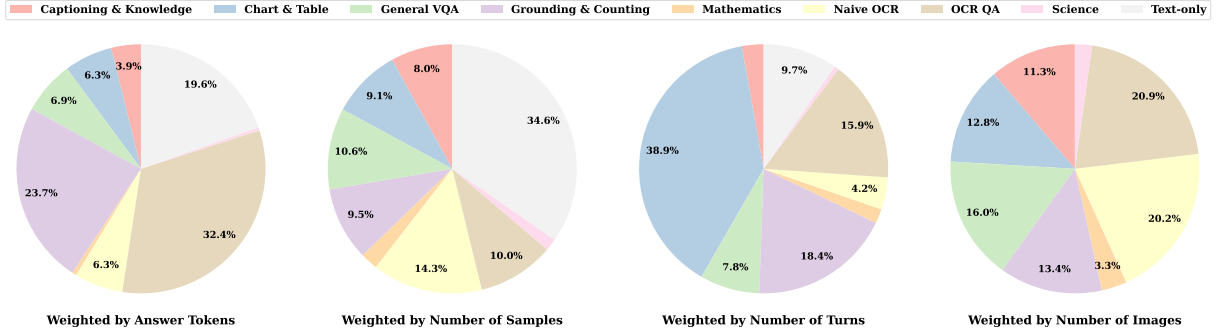


Figure 3 | Category distribution. Share of samples across the nine categories. FineVision provides a good baseline mixture, which could be further tuned via up- and downsampling and in correlation with the quality ratings.

3 Exploring FineVision

We characterize FineVision along three key axes: category composition, turn quality, and visual diversity.

3.1 Category Composition

We categorize every FineVision subset into nine distinct categories, following Li et al. (2025b): **Captioning & Knowledge**, **Chart & Table**, **General VQA**, **Grounding & Counting**, **Mathematics**, **Naive OCR**, **OCR QA**, **Science** and **Text-only**. We analyze the resulting category composition along multiple axes: number of images, samples, turns, and answer tokens (see Fig. 3). Samples from **Chart & Table** usually lend themselves well to multi-turn conversations, since multiple similar questions can be asked for a single Chart. Samples from **OCR QA** tend to have longer answers, since they aim at detailed document understanding, which are rarely answered with a short sentence. For in-depth statistics on token length, conversation turns, and image resolution by category, see Appendix A.6.

3.2 Analysis of Characteristic Axes

We characterize every training turn by scoring it from 1-5 with LLM/VLM-as-a-judge (Qwen3-32B for text-only criteria and Qwen2.5VL-32B-Instruct for image-conditioned criteria, served locally via vLLM) along four **characteristic axes**: *Formatting*, *Relevance*, *Visual Dependency*, and *Image-Question Correspondence* (see Appendix A.2 for the full prompts). Fig. 4 shows that *Relevance* is uniformly high across categories, with more than 85% of the turns scoring 4 or 5, and *Formatting* scores are high overall, with 97.2% of the turns scoring 4 or 5, peaking for **Grounding**. These two text-based axes confirm that FineVision pairs well-formed questions with answers that stay on-topic.

As can be seen in Fig. 4, the vision-centric axes distinguish task nature most clearly. **Captioning** and **General VQA** achieve high scores on both *Visual Dependency* and *Formatting/Relevance*, alongside low scores in *Image-Question Correspondence*. **Naive OCR** also has high *Visual Dependency*, but with lower scores on the other axes. By contrast, **Mathematics** shows a different profile, exhibiting lower scores across all four axes. **Chart & Table** is defined by high *Image-Question Correspondence* and *Formatting/Relevance*, but lower *Visual Dependency*, mixing low-dependency lookups with higher-dependency integrative cases (e.g., comparing trends across multiple series rather than retrieving a single value), consistent with the variability between reading values and reasoning across trends.

The cross-axis patterns further clarify these roles (see Fig. 5). The two vision axes—*Visual Dependency* and *Image-Question Correspondence*—are inversely correlated, indicating that tasks requiring the image for an answer often differ from those where the question directly corresponds to image content. Conversely, *Formatting* and *Relevance* trend together but remain partly orthogonal to the vision-centric axes. This is evident when comparing **Grounding**, which scores highly on text-based axes, against **Naive OCR**, which is highly visually dependent but scores lower on *Formatting* and *Relevance*. We release per-turn scores to support analysis and reweighting; in our experiments, preserving breadth rather than aggressive filtering yields the best downstream generalization (see Appendix A.4).

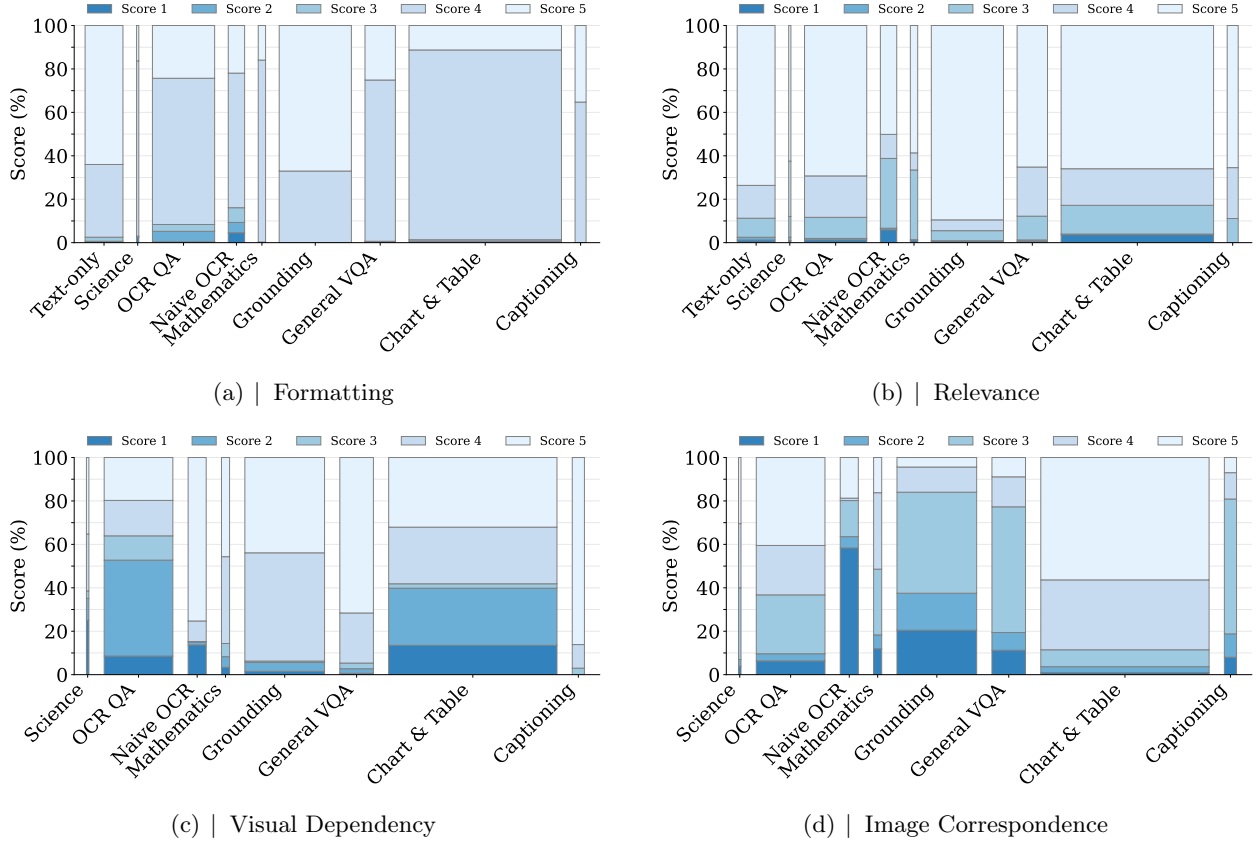


Figure 4 | Quality rating distributions by category. Score distributions across the four quality axes (Top left: Formatting, Top right: Relevance, Bottom left: Visual Dependency, Bottom right: Image-Question Correspondence) broken down by dataset category. Category width corresponds to the number of turns.

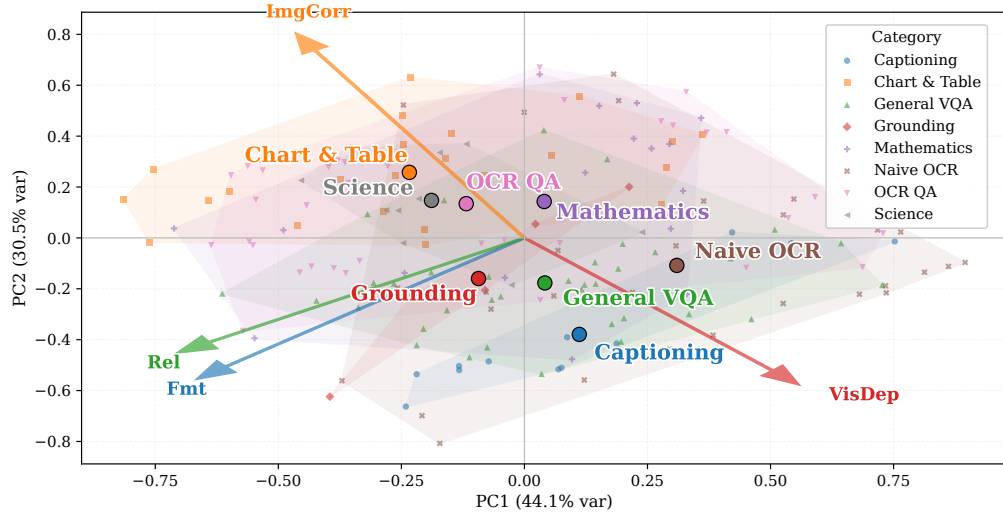


Figure 5 | Dataset Characterization Along Four Axes We apply per-dataset PCA over the different characteristic scores. From the analysis, it appears *Formatting* and *Relevance* are highly correlated, while *Visual Dependency* and *Image-Question Correspondence* are strongly inversely correlated. *Grounding* attains the highest *Formatting/Relevance*, while *Chart & Table* attains high *Image-Question Correspondence*. *Captioning* and *General VQA* show high *Visual Dependency* combined with strong *Formatting/Relevance*. In contrast, *Naive OCR* exhibits high *Visual Dependency* but lower scores on *Formatting/Relevance*. Arrows indicate variable loadings; points are dataset centroids with covariance ellipses per category.

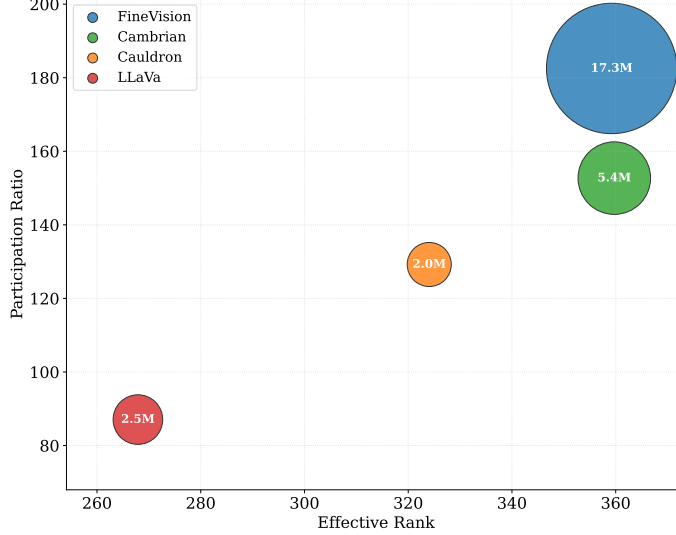


Figure 6 | Visual diversity analysis. While FineVision and Cambrian show similarly high conceptual breadth (effective rank), FineVision exhibits superior conceptual balance (participation ratio). Higher values are better for both axes; both axes are linear and dimensionless. Marker size corresponds to the number of images in each dataset. For intuition, a dataset with high effective rank but low participation ratio might cover many animal species but be numerically dominated by a few (e.g., cats/dogs).

3.3 Visual Diversity

We analyze visual diversity using covariance-spectrum statistics of self-supervised copy-detection (SSCD) embeddings (Pizzi et al., 2022) (same pipeline as for deduplication), computed per dataset without subsampling. SSCD descriptors are optimized to distinguish near-duplicates and, via entropy regularization, promote uniform occupancy of the embedding space, making distances more comparable across regions and suitable for diversity measurement. Let λ_i be the eigenvalues of the embedding covariance; the images are normalized and resized the same way before processing, and the embeddings are not specifically centered. We report two complementary measures:

- *Effective Rank* $r_{\text{eff}} = \exp(H(p))$ with $p_i = \lambda_i / \sum_j \lambda_j$ and $H(p) = -\sum_i p_i \log p_i$ (equivalent to the Vendi Score (Friedman and Dieng, 2023)). Higher values indicate that the variance is spread across more dimensions, signifying a greater conceptual breadth.
- *Participation Ratio* $PR = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$. Higher values indicate that the variance is distributed more uniformly across dimensions, indicating a more balanced data set.

As shown in Fig. 6, these metrics reveal a clear separation of the datasets. FineVision and Cambrian occupy a high-diversity tier, demonstrating significantly greater conceptual breadth (effective rank) than Cauldron and LLaVa, whose narrower scope may limit the world knowledge of models trained on them.

However, the most crucial insight emerges from the high-diversity tier. Although both FineVision and Cambrian exhibit a similarly high effective rank – indicating they cover a comparably broad range of visual concepts – FineVision possesses a substantially higher participation ratio. This distinction is key, as it shows that FineVision’s conceptual coverage is not only broad but also significantly more uniform. Its variance is more evenly distributed between concepts, providing a stronger foundation for training models that are robust and generalize well. We compute these metrics on the full datasets without subsampling and therefore do not report confidence intervals; given large size differences, naive bootstrapping would be misleading. Covariances are computed in a numerically stable way (e.g., via Welford’s algorithm).

Finally, dataset size (marker size) alone does not explain diversity; curation strategy is equally critical. FineVision’s success lies in achieving both massive scale and best-in-class conceptual balance. Exact dataset sizes are reported in Table 1.

Dataset	Size Statistics				Diversity Summary		
	Images	Samples	Turns	Ans. Tok.	Eff. Rank	Part. Ratio	
Cauldron	2.0M	1.8M	27.8M	0.3B	324.05	129.22	
LLaVA	2.5M	3.9M	9.1M	1.0B	267.89	87.05	
Cambrian	5.4M	7.1M	12.2M	0.8B	359.73	152.70	
FineVision	17.3M	24.3M	88.9M	9.5B	359.22	182.52	

Table 1 | Comparison of dataset size and diversity. Size metrics (images, samples, turns, answer tokens) and diversity metrics (effective rank, participation ratio). FineVision is both substantially bigger and more diverse than the baselines.

4 Experiments and Results

We conduct a series of experiments to validate the effectiveness of FineVision. We establish the experimental setup, then present our main results comparing FineVision to existing datasets and finally evaluate novel capabilities.

4.1 Experimental Setup

Model and Training. For all experiments, we train a 460M-parameter SmolVLM (Marafioti et al., 2025) using the nanoVLM framework (Wiedmann et al., 2025). The architecture consists of a **SmolLM2-360M-Instruct** (Al-lal et al., 2025) text backbone and a **SigLIP2-Base-512** (Tschannen et al., 2025) vision encoder. Unless otherwise specified, we employ a single-stage training protocol for 20,000 steps with an effective batch size of 512, which takes approximately 20 hours on 32 H100 GPUs. With sequence packing to the max length of 8192, this covers more than one effective epoch over the FineVision dataset.

Baselines. We compare FineVision against three prominent open-source datasets: The Cauldron (Laurençon et al., 2024), LLaVA-OneVision (Li et al., 2024), and Cambrian-7M (Tong et al., 2024). Table 1 summarizes their respective scales and diversity scores.

Evaluation. We use the **lmms-eval** framework (Zhang et al., 2024c) to evaluate models on a diverse suite of 11 benchmarks, comprising AI2D, ChartQA, DocVQA, InfoVQA, MME, MMMU, ScienceQA, MMStar, OCRBench, TextVQA and SEED-Bench.

4.2 Main Results

Comparison with Existing Datasets. Models trained on FineVision significantly outperform those trained on other open-source datasets. As shown in Figure 7, the FineVision-trained model achieves the highest average performance across all 11 benchmarks (left). While it initially lags behind during the first few thousand steps – likely due to the inclusion of novel tasks not present in the baselines – it surpasses all other models after approximately one epoch of training, demonstrating superior generalization. By the end of training, FineVision yields an average absolute score improvement of 12.7 percentage points (pp) over The Cauldron, 5.1 pp over Cambrian-7M, and 14.3 pp over LLaVA-OneVision.

Impact of Test Data Contamination. We investigated test set leakage by processing all datasets through the same pipeline described in Section 2.4 and found that the baseline datasets contain between 2.15–3.05% of images that are also present in common evaluation benchmarks. FineVision has a contamination rate of only 1.02%. When we retrained all models on decontaminated versions of their respective datasets, the performance of baseline models dropped by 2.7–3.7 pp, while the FineVision model’s performance dropped by only 1.6 pp (for full details see Appendix Tab. 4).

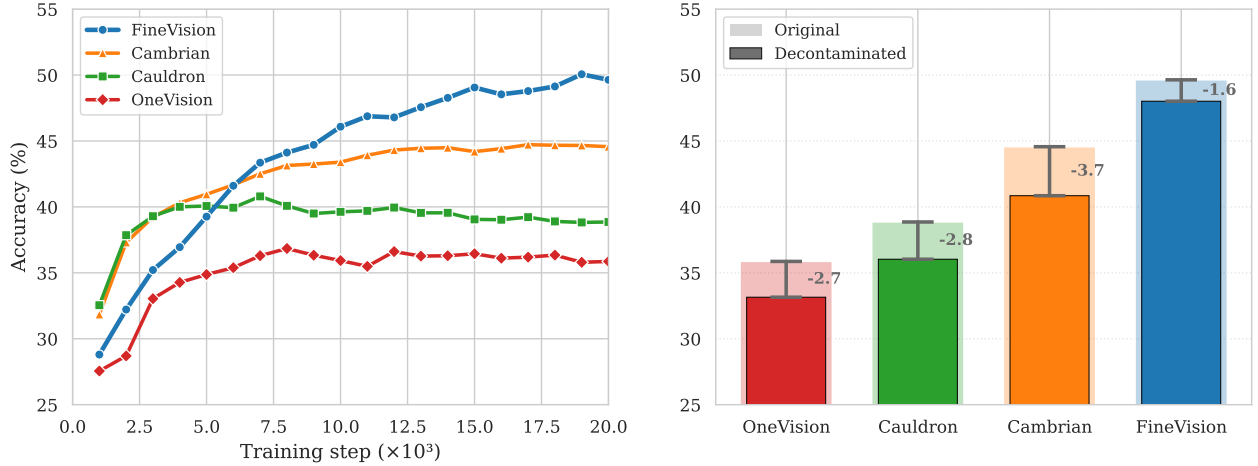


Figure 7 | Training dynamics on original and decontaminated datasets. **Left:** mean normalized performance (%) across 11 evaluation benchmarks (higher is better), with the training step shown in thousands ($\times 10^3$). Each benchmark score is min-max normalized to [0,100] and averaged per evaluation step; the model trained on FineVision (blue) leads throughout the second half of training and attains the best final score. **Right:** comparison of the final performance between the original data and after decontamination. FineVision exhibits the smallest drop at the end of training with 1.6 pp, whereas baselines degrade by roughly 2.7–3.7 pp, indicating that FineVision’s gains are not explained by contamination.

New GUI capabilities. FineVision contains substantial amounts of GUI/agent data, which represents an important new capability for VLMs. In addition, tasks measuring the performance in this domain are not available in standard evaluation frameworks yet, hindering the widespread tracking of this capability. We compare the same 460M model trained on FineVision (FV-0.5B) with an architecturally equivalent SmolVLM2 in two sizes (Smol-2B³ and Smol-0.5B⁴) on **Screenspot-V2** (Wu et al., 2024b) and **Screenspot-Pro** (Li et al., 2025a). Since these benchmarks are quite challenging for small open models, we report both the performance of the base models as well as after fine-tuning on one epoch of the **aguis-stage-1** subset, which is also part of FineVision. Most small models fail to solve any task at their base stage, and after fine-tuning FineVision-trained models achieve results on par with an architecturally equivalent model 4x its size.

	Base Models			FineTuned		
	Smol-2B	Smol-0.5B	FV-0.5B	Smol-2B	Smol-0.5B	FV-0.5B
ScreenSpot-Pro	0.00	0.00	0.00	0.07	0.01	0.06
ScreenSpot-V2	0.00	0.00	0.20	0.41	0.24	0.48

Table 2 | Comparison of model performance on ScreenSpot. While this benchmark is challenging for small open models, the FineVision-trained model shows strong performance and achieves comparable results to an architecturally equivalent model 4x its size.

³<https://huggingface.co/HuggingFaceTB/SmolVLM2-2.2B-Instruct>

⁴<https://huggingface.co/HuggingFaceTB/SmolVLM2-500M-Video-Instruct>

5 Related Work

Large-Scale Multimodal Data Generation Pipelines (New Data Creation). This line of work creates new large-scale multimodal datasets via synthetic generation or multi-expert fusion to overcome the scalability limits of human annotation. Early pipelines like *LLaVA-Instruct-150K* (Liu et al., 2023d) (158K image-instruction pairs over ~ 118 K COCO images) demonstrated GPT-4-generated multimodal instructions guided by BLIP/CLIP-style embeddings. Specialized generation then scaled in multiple directions: *DenseFusion-1M* (Li et al., 2024b) (1.06M pairs from LAION-5B) uses a two-stage perceptual-fusion pipeline that integrates object detectors, OCR, and depth estimators with a multimodal model, plus error filtering, to produce hyper-detailed single-paragraph captions; *ShareGPT4V* (Chen et al., 2024b) develops a seed (100K) $\rightarrow$ expansion (1.2M) recipe using GPT-4V followed by ShareCaptioner with length/content quality filters; and *WebSight* (Laurençon et al., 2024) synthesizes ~ 2 M webpage screenshots from LLM-generated HTML/CSS (Tailwind), applying rendering/quality filters and removing unsupported/noisy pages to create perfectly aligned UI-image-code pairs. Document-centric pipelines push reading supervision: *DocVLM* (Nacson et al., 2024) instruments high-resolution documents with OCR for efficient reading, while large meta-collections such as *Docmatix* (Laurençon et al., 2024) (~ 9.5 M QA over ~ 2.4 M document images) filter $\sim 15\%$ hallucinated or unanswerable QAs. Fine-grained generators (e.g., *LVIS-Instruct4V* (Wang et al., 2023a)) and region-level prompting (e.g., *ViP-LLaVA-Instruct* (Cai et al., 2024)) emphasize localized grounding, and web-scale interleaved corpora (e.g., *MMC4* (Zhu et al., 2023), *OBELICS* (Laurençon et al., 2023)) complement instruction data via heavy filtering of raw web documents. Common safeguards across pipelines include expert-fusion signals, rendering/consistency checks, and targeted content/length filters, which together yield denser and more structured supervision than legacy caption/VQA corpora.

Meta-Datasets for Multimodal Instruction Tuning. The development of large-scale multimodal instruction datasets has rapidly evolved to address the growing demands of vision-language models. Early efforts like *MultiInstruct* (Xu et al., 2023) pioneered the field with ~ 510 K fully human-annotated instances across 62 diverse tasks, establishing high-quality instruction-following as a priority. *InstructBLIP* (Dai et al., 2023) scaled this approach to ~ 1.6 M instances by aggregating ~ 12 existing datasets through templated conversion, trading manual curation for breadth. The field matured in 2024 with several ambitious collections: *Vision-FLAN* (Xu et al., 2024) brought rigorous human curation to ~ 1.66 M instances across 187 tasks from 101 datasets, emphasizing expert-written instructions; *Cambrian-10M* (Tong et al., 2024) pushed scale boundaries with ~ 10 M images and introduced a balanced 7M subset to address quality-quantity trade-offs; and *The Cauldron* (Laurençon et al., 2024) unified 50+ datasets into ~ 30 M multi-turn dialogues for Idefics2, applying targeted test-set decontamination rather than broad internal de-duplication. *LLaVA-OneVision* (Li et al., 2024) carefully curated ~ 3.9 M instruction-response pairs (~ 1.2 M images), extending image SFT to multi-image reasoning and video understanding, with strengthened document/OCR and multilingual coverage. Most recently, *MAmmoTH-VL-Instruct* (Guo et al., 2025) demonstrated the potential of fully synthetic pipelines, using open-source models plus filtering to generate ~ 12 M rational-augmented pairs with detailed reasoning chains, and scaled chain-of-thought supervision for multimodal tasks. Other recent works (Li et al., 2025b; Cho et al., 2025) also cite the utilization of an order of 200 datasets. Our work, FineVision, addresses these limitations by unifying 185 open sources into a 24M-sample corpus via a semi-automated, human-in-the-loop pipeline that preserves task structure and conversational formatting, applies rigorous intra-/cross-source de-duplication and decontamination against 66 public benchmarks, and extends coverage to agentic/GUI tasks with a unified action space, yielding state-of-the-art results among open-data mixtures.

GUI and Embodied Vision Datasets. A newer frontier links perception to action – models acting in GUIs or embodied environments. *OS-Atlas* (Zhiyong et al., 2024) introduced a cross-platform GUI corpus with over 2.3M screenshots and 13M GUI elements spanning web, desktop, and mobile interfaces, and trained a 7B LVLM with a unified function-call API for UI manipulation. *ShowUI* (Lin et al., 2025) presents a vision-language-action model that treats GUI automation as sequence modeling; a lightweight 2B model trained on 256k high-quality interaction steps achieves strong zero-shot grounding. Complementary efforts target robust GUI grounding and control, including *UIShift* (Gao et al., 2025) and *GUI-Actor* (Wu et al., 2025). Most GUI agents adopt a unified action space, predicting structured actions (e.g., clicks, typing) as next tokens; cross-platform ambiguities and the limited scale of high-quality interaction data remain open

challenges.

6 Conclusion

We introduced FINEVISION, a large-scale, open, and rigorously curated dataset for training vision–language models. Through a semi-automated, human-in-the-loop pipeline that unifies over 200 public sources into a standardized conversational schema, we deliver high-quality supervision spanning captions, VQA, document understanding, OCR, grounding, and GUI interaction. Our pipeline integrates systematic cleaning, near-duplicate control, and benchmark decontamination using SSCD-based matching, enabling reproducible and hygienic training data.

Empirically, models trained on FINEVISION consistently outperform those trained on existing open datasets across a broad suite of benchmarks, and the gains persist after test-set decontamination. Beyond aggregate scores, FINEVISION broadens capabilities – particularly for GUI/agent settings via a unified action space – suggesting that scale paired with targeted diversity matters for generalization.

We release the dataset, conversion recipes, de-duplication tools, and precomputed embeddings to foster transparent, repeatable research. While our curation reduces leakage and noisy supervision, limitations remain: residual overlaps may persist, long-context and multi-document reasoning are still challenging, and community benchmarks for GUI control are not integrated into the standard training stack. We adhere to source licensing and apply safety-oriented filters; future work will strengthen audits for licensing provenance, privacy, and bias. We view FINEVISION as a foundation and invite the community to extend it to video, richer multilingual coverage, longer-context reasoning, and stronger human evaluation protocols, further closing the gap between open and proprietary VLM training data.

References

- Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084, 2019. <https://ojs.aaai.org/index.php/AAAI/article/view/4815>. Issue: 01.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgrén, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. <https://arxiv.org/abs/2502.02737>.
- Jonas Belouadi, Anne Lauscher, and Steffen Eger. AutomaTikZ: Text-guided synthesis of scientific vector graphics with TikZ, 2023. <http://arxiv.org/abs/2310.00367>.
- Shreyanshu Bhushan and Minh Lee. Block diagram-to-text: Understanding block diagram images by generating natural language descriptors. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, pages 153–168, 2022. <https://aclanthology.org/2022.findings-aacl.15/>.
- Tarun Bisht. iamtarun/python_code_instructions_18k_alpaca via datasets at hugging face, 2024. https://huggingface.co/datasets/iamtarun/python_code_instructions_18k_alpaca.
- Ali Furkan Biten, Ruben Tito, Andres Mafra, Lluís Gomez, Marçal Rusinol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. http://openaccess.thecvf.com/content_ICCV_2019/html/Biten_Scene_Text_Visual_Question_Answering_ICCV_2019_paper.html.
- Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.
- Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pages 1511–1520, 2022. <https://aclanthology.org/2022.coling-1.130/>.

- Jimmy Carter. jimmycarter/textocr-gpt4v via datasets at hugging face, 2024. <https://huggingface.co/datasets/jimmycarter/textocr-gpt4v>.
- Sunguk Cha, Jusung Lee, Younghyun Lee, and Cheoljong Yang. Visually dehallucinative instruction generation: Know what you don’t know, 2024. <http://arxiv.org/abs/2402.09717>.
- Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. MapQA: A dataset for question answering on choropleth maps, 2022. <http://arxiv.org/abs/2211.08545>.
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. ALLaVA: Harnessing GPT4v-synthesized data for lite vision-language models, 2024a. <http://arxiv.org/abs/2402.11684>.
- Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression, 2022a. <https://arxiv.org/abs/2212.02746>.
- 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. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15075, pages 370–387. Springer Nature Switzerland, 2024b. ISBN 978-3-031-72642-2 978-3-031-72643-9. doi: 10.1007/978-3-031-72643-9_22. https://link.springer.com/10.1007/978-3-031-72643-9_22. Series Title: Lecture Notes in Computer Science.
- Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset, 2023. <http://arxiv.org/abs/2305.12524>.
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data, 2022b. <http://arxiv.org/abs/2109.00122>.
- Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. HiTab: A hierarchical table dataset for question answering and natural language generation, 2022. <http://arxiv.org/abs/2108.06712>.
- Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, and Errui Ding. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. <https://ieeexplore.ieee.org/abstract/document/8978157/>.
- Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Brian Werness, Ari Morcos, and Lingpeng Xie. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. <https://arxiv.org/abs/2504.13180>.
- LLM-Red-Team Contributors. emo-visualdata: Emotion and visual data analysis project, 2024. <https://github.com/LLM-Red-Team/emo-visual-data>.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.
- Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, Jose M. Saavedra, David Contreras, Juan Manuel Barrios, and Luiz S. Oliveira. ICFHR 2014 competition on handwritten digit string recognition in challenging datasets (HDSRC 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 779–784. IEEE, 2014. <https://ieeexplore.ieee.org/abstract/document/6981115/>.
- Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. Vqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 585–601. Springer, 2023.
- EvolvingLMs-Lab. Lmms-eval: Comprehensive evaluation suite for lvlms. <https://github.com/EvolvingLMs-Lab/lmms-eval>, 2024.
- FastJobs. Fastjobs/visual_emotional_analysis via datasets at hugging face, 2024. https://huggingface.co/datasets/FastJobs/Visual_Emoional_Analysis.
- FLOCK4H. flytech/python-codes-25k via datasets at hugging face, 2024. <https://huggingface.co/datasets/flytech/python-codes-25k>.

- Inked Forms. `ift/handwriting_forms` via datasets at hugging face, 2024. https://huggingface.co/datasets/ift/handwriting_forms.
- Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.
- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-LLaVA: Solving geometric problem with multi-modal large language model, 2023. <http://arxiv.org/abs/2312.11370>.
- Longxi Gao, Li Zhang, and Mengwei Xu. Uishift: Enhancing vlm-based gui agents through self-supervised reinforcement learning. *arXiv preprint arXiv:2505.12493*, 2025. <https://arxiv.org/abs/2505.12493>.
- Philippe Gervais, Anastasiia Fadeeva, and Andrii Maksai. MathWriting: A dataset for handwritten mathematical expression recognition. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 5459–5469. ACM, 2025. ISBN 979-8-4007-1454-2. doi: 10.1145/3711896.3737436. <https://dl.acm.org/doi/10.1145/3711896.3737436>.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. http://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html.
- Lianmin Guo, Yujie Lu, Weihao Wang, Shengding Huang, Yujia Chen, Yanzhou Lin, Wenxuan Xia, Haozhe Wu, Yingfa Han, Wentao Xue, Zhiyuan Liu, and Maosong Sun. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2025. <https://arxiv.org/abs/2412.05237>.
- Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. http://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering, 2020. <http://arxiv.org/abs/2003.10286>.
- Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Srinivas Sunkara, Victor Carbune, Jason Lin, Maria Wang, Yun Zhu, and Jindong Chen. ScreenQA: Large-scale question-answer pairs over mobile app screenshots, 2022. <http://arxiv.org/abs/2209.08199>.
- Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. <https://ieeexplore.ieee.org/abstract/document/8977955/>.
- Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019. <https://ieeexplore.ieee.org/abstract/document/8892998/>.
- Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images, 2018. <http://arxiv.org/abs/1808.10584>.
- Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhui Chen. Visualwebinstruct: Scaling up multimodal instruction data through web search, 2025.
- Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. http://openaccess.thecvf.com/content_cvpr_2018/html/Kafle_DVQA_Understanding_Data_CVPR_2018_paper.html.
- Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning, 2017. <http://arxiv.org/abs/1710.07300>.
- Yasindu Kamizuru. `Kamizuru00/diagram_image_to_text` via datasets at hugging face, 2024. https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text.
- Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization, 2022. <http://arxiv.org/abs/2203.06486>.

- Hazal Karakus. hazal-karakus/mscoco-controlnet-canny-less-colors via datasets at hugging face, 2024. <https://huggingface.co/datasets/hazal-karakus/mscoco-controlnet-canny-less-colors>.
- Mehran Kazemi, Hamidreza Alvani, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. GeomVerse: A systematic evaluation of large models for geometric reasoning, 2023. <http://arxiv.org/abs/2312.12241>.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. <https://arxiv.org/abs/1603.07396>.
- Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017. http://openaccess.thecvf.com/content_cvpr_2017/html/Kembhavi_Are_You_Smarter_CVPR_2017_paper.html.
- Kerem. keremberke/indoor-scene-classification via datasets at hugging face, 2024. <https://huggingface.co/datasets/keremberke/indoor-scene-classification>.
- 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. *Advances in neural information processing systems*, 33:2611–2624, 2020. <https://proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html>.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13688, pages 498–517. Springer Nature Switzerland, 2022. ISBN 978-3-031-19814-4 978-3-031-19815-1. doi: 10.1007/978-3-031-19815-1_29. https://link.springer.com/10.1007/978-3-031-19815-1_29. Series Title: Lecture Notes in Computer Science.
- Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. Textstylebrush: transfer of text aesthetics from a single example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9122–9134, 2023. <https://ieeexplore.ieee.org/abstract/document/10027471/>. Publisher: IEEE.
- Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-DPO: Step-wise preference optimization for long-chain reasoning of LLMs, 2024. <http://arxiv.org/abs/2406.18629>.
- LAION. laion/gpt4v-dataset via datasets at hugging face, 2023. <https://huggingface.co/datasets/laion/gpt4v-dataset>.
- Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. <https://www.nature.com/articles/sdata2018251>. Publisher: Nature Publishing Group.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 87874–87907. Curran Associates, Inc., 2024. https://proceedings.neurips.cc/paper_files/paper/2024/file/a03037317560b8c5f2fb4b6466d4c439-Paper-Conference.pdf.
- Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.
- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *NeurIPS 2024 Workshop RBFM*, 2024.
- Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into HTML code with the WebSight dataset, 2024. <http://arxiv.org/abs/2403.09029>.
- 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. *arXiv preprint arXiv:2408.03326*, 2024. doi: 10.48550/arXiv.2408.03326. <https://arxiv.org/abs/2408.03326>.
- Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. [<https://huggingface.co/AI-MO/NuminaMath-CoT>] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

- 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. *arXiv preprint arXiv:2504.07981*, 2025a.
- Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models, 2024a. <http://arxiv.org/abs/2403.00231>.
- Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Lingyu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *Advances in Neural Information Processing Systems*, 37:18535–18556, 2024b. https://proceedings.neurips.cc/paper_files/paper/2024/hash/20ffc2b42c7de4a1960cfdadf305bbe2-Abstract-Datasets_and_Benchmarks_Track.html.
- Zekun Li, Yijun Lin, Yao-Yi Chiang, Jerod Weinman, Solenn Tual, Joseph Chazalon, Julien Perret, Bertrand Duménieu, and Nathalie Abadie. ICDAR 2024 competition on historical map text detection, recognition, and linking. In Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng, editors, *Document Analysis and Recognition - ICDAR 2024*, volume 14809, pages 363–380. Springer Nature Switzerland, 2024c. ISBN 978-3-031-70551-9 978-3-031-70552-6. doi: 10.1007/978-3-031-70552-6_22. https://link.springer.com/10.1007/978-3-031-70552-6_22. Series Title: Lecture Notes in Computer Science.
- Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Chen Qian, Zhengkai Song, Qihang Xu, Chuzhao Guo, Peize Sun, Kai Chen, Shuicheng Yan, Ping Luo, and Kaipeng Zhang. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025b. <https://arxiv.org/abs/2501.14818>.
- Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L. Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973, 2023. http://openaccess.thecvf.com/content/CVPR2023/html/Li_Super-CLEVR_A_Virtual_Benchmark_To_Diagnose_Domain_Robustness_in_Visual_CVPR_2023_paper.html.
- Wing Lian, Bley Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces. <https://huggingface.co/datasets/Open-Orca/OpenOrca>, 2023.
- Kevin Qinghong Lin, Linjie Zhou, Rohit Girdhar, Yiqi Zhu, Renrui Zhang, Yan Yan, Hao Fang, Wei Li, Chunyuan Xiao, Zicheng Zhang, and Chunyuan Li. Showui: One vision-language-action model for gui visual agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. https://openaccess.thecvf.com/content/CVPR2025/papers/Lin_ShowUI_One_Vision-Language-Action_Model_for_GUI_Visual_Agent_CVPR_2025_paper.pdf.
- Adam Dahlgren Lindström and Savitha Sam Abraham. CLEVR-math: A dataset for compositional language, visual and mathematical reasoning, 2022. <http://arxiv.org/abs/2208.05358>.
- Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a. https://direct.mit.edu/tac1/article/doi/10.1162/tac1_a_00566/116470. Publisher: MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2023b. <http://arxiv.org/abs/2306.14565>.
- Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. MMC: Advancing multimodal chart understanding with large-scale instruction tuning, 2023c. <http://arxiv.org/abs/2311.10774>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023d. <https://arxiv.org/abs/2304.08485>. arXiv:2304.08485.
- LMMS-Lab. lmms-lab/LLaVA-OneVision-data via datasets at hugging face, 2025. <https://huggingface.co/datasets/lmms-lab/LLaVA-OneVision-Data>.
- LooksJuicy. Looksjuicy/ruozhiba via datasets at hugging face, 2024. <https://huggingface.co/datasets/LooksJuicy/ruozhiba>.
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning, 2021a. <http://arxiv.org/abs/2105.04165>.

- Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning, 2021b. <http://arxiv.org/abs/2110.13214>.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.
- Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning, 2023. <http://arxiv.org/abs/2209.14610>.
- Yujie Lu, Dongfu Jiang, Wenhui Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *Advances in Neural Information Processing Systems*, 37:48224–48255, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/563991b5c8b45fe75bea42db738223b2-Abstract-Datasets_and_Benchmarks_Track.html.
- Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, Heng Tao Shen, Yunshui Li, Xiaobo Xia, Fei Huang, Jingkuan Song, and Yongbin Li. MMEvol: Empowering multimodal large language models with evol-instruct, 2024. <http://arxiv.org/abs/2409.05840>.
- Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. In *Conference on Language Modeling*, 2025.
- U.-V. Marti and H. Bunke. The IAM-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002. ISSN 1433-2833, 1433-2825. doi: 10.1007/s100320200071. <http://link.springer.com/10.1007/s100320200071>.
- 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. <http://arxiv.org/abs/2203.10244>.
- Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning, 2023. <http://arxiv.org/abs/2305.14761>.
- Minesh Mathew, Lluís Gomez, Dimosthenis Karatzas, and C. V. Jawahar. Asking questions on handwritten document collections. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3):235–249, 2021a. ISSN 1433-2833, 1433-2825. doi: 10.1007/s10032-021-00383-3. <https://link.springer.com/10.1007/s10032-021-00383-3>.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021b. http://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. http://openaccess.thecvf.com/content/WACV2022/html/Mathew_InfographicVQA_WACV_2022_paper.html.
- Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1527–1536, 2020. http://openaccess.thecvf.com/content_WACV_2020/html/Methani_PlotQA_Reasoning_over_Scientific_Plots_WACV_2020_paper.html.
- Minyang. mychen76/invoices-and-receipts_ocr_v1 via datasets at hugging face, 2024. https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1.
- Anand Mishra, Karteek Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012. <https://inria.hal.science/hal-00818183/>.
- Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. <https://ieeexplore.ieee.org/abstract/document/8978122/>.
- Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of SLMs in grade school math, 2024. <http://arxiv.org/abs/2402.14830>.

- Harold Mouchere, Christian Viard-Gaudin, Richard Zanibbi, Utpal Garain, Dae Hwan Kim, and Jin Hyung Kim. Icdar 2013 crohme: Third international competition on recognition of online handwritten mathematical expressions. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1428–1432. IEEE, 2013. <https://ieeexplore.ieee.org/abstract/document/6628849/>.
- Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Shai Mazor, and Ron Litman. Docvln: Make your vlm an efficient reader. *arXiv preprint arXiv:2412.08746*, 2024. <https://arxiv.org/abs/2412.08746>.
- Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. YesBut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models, 2024. <http://arxiv.org/abs/2409.13592>.
- Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A. Said Gurbuz, Michele Dolfi, Miquel Farré, and Peter W. J. Staar. SmolDocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion, 2025. <http://arxiv.org/abs/2503.11576>.
- OleehyO. OleehyO/latex-formulas via datasets at hugging face, 2024. <https://huggingface.co/datasets/OleehyO/latex-formulas>.
- OpenGVLab. ShareGPT-4o, 2024. <https://sharegpt4o.github.io/>.
- Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions, 2024. <http://arxiv.org/abs/2406.07502>.
- Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16532–16541, 2022. https://openaccess.thecvf.com/content/CVPR2022/html/Pizzi_A_Self-Supervised_Descriptor_for_Image_Copy_Detection_CVPR_2022_paper.html.
- Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12350, pages 647–664. Springer International Publishing, 2020. ISBN 978-3-030-58557-0 978-3-030-58558-7. doi: 10.1007/978-3-030-58558-7_38. https://link.springer.com/10.1007/978-3-030-58558-7_38. Series Title: Lecture Notes in Computer Science.
- Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M Alabdulmohsin. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. *Advances in Neural Information Processing Systems*, 37:106474–106496, 2024.
- Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmOCR: Unlocking trillions of tokens in PDFs with vision language models, 2025. <http://arxiv.org/abs/2502.18443>.
- Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: workshop on multimodal fact checking and hate speech detection, CEUR*, volume 17, 2022.
- Yaseen Razeghi et al. Impact of pretraining term frequencies on few-shot reasoning, 2022.
- Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015. <https://proceedings.neurips.cc/paper/2015/hash/831c2f88a604a07ca94314b56a4921b8-Abstract.html>.
- Parsa Samadnejad. Captcha dataset, 2024. <https://www.kaggle.com/datasets/parsasam/captcha-dataset>.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13668, pages 146–162. Springer Nature Switzerland, 2022. ISBN 978-3-031-20073-1 978-3-031-20074-8. doi: 10.1007/978-3-031-20074-8_9. https://link.springer.com/10.1007/978-3-031-20074-8_9. Series Title: Lecture Notes in Computer Science.
- Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015. <https://aclanthology.org/D15-1171.pdf>.

- Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438. IEEE, 2019. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00852. <https://ieeexplore.ieee.org/document/9009553/>.
- Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models, 2024. <http://arxiv.org/abs/2406.17294>.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning, 2021. <http://arxiv.org/abs/2010.03768>.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A dataset for image captioning with reading comprehension. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12347, pages 742–758. Springer International Publishing, 2020. ISBN 978-3-030-58535-8 978-3-030-58536-5. doi: 10.1007/978-3-030-58536-5_44. https://link.springer.com/10.1007/978-3-030-58536-5_44. Series Title: Lecture Notes in Computer Science.
- Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15110, pages 256–274. Springer Nature Switzerland, 2024. ISBN 978-3-031-72942-3 978-3-031-72943-0. doi: 10.1007/978-3-031-72943-0_15. https://link.springer.com/10.1007/978-3-031-72943-0_15. Series Title: Lecture Notes in Computer Science.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.
- Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017.
- Hamed Rahimi Sujei AI, Allaa Boutaleb. Sujei-finance-qa-vision-100k: A large-scale dataset for financial document vqa, 2024. <https://huggingface.co/datasets/sujei-ai/Sujei-Finance-QA-Vision-100k>.
- TAL. Tal open dataset, 2023. <https://ai.100tal.com/dataset>.
- Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021. <https://ojs.aaai.org/index.php/AAAI/article/view/17635>. Issue: 15.
- Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645, 2023. <https://ojs.aaai.org/index.php/AAAI/article/view/26598>. Issue: 11.
- Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. VisText: A benchmark for semantically rich chart captioning, 2023. <http://arxiv.org/abs/2307.05356>.
- Teknum. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. <https://huggingface.co/datasets/teknum/OpenHermes-2.5>.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/file/9ee3a664ccfeabc0da16ac6f1f1cfe59-Paper-Conference.pdf.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data, 2024. <http://arxiv.org/abs/2410.01560>.
- Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyers, 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. <https://arxiv.org/abs/2502.14786>.
- Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision LLMs, 2023. <http://arxiv.org/abs/2311.16101>.
- Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. COCO-text: Dataset and benchmark for text detection and recognition in natural images, 2016. <http://arxiv.org/abs/1601.07140>.
- VQAonBD. Vqaonbd dataset, 2023. <https://ilocr.iiit.ac.in/vqabd/dataset.html>.
- Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510, 2021.
- Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023a.
- Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting GPT-4v for better visual instruction tuning, 2023b. <http://arxiv.org/abs/2311.07574>.
- Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020. http://openaccess.thecvf.com/content_CVPR_2020/html/Wang_On_the_General_Value_of_Evidence_and_Bilingual_Scene-Text_Visual_CVPR_2020_paper.html.
- Yizhong Wang et al. Rethinking the role of llms as evaluators, 2023c.
- Chris Wendler. wendlerc/RenderedText via datasets at hugging face, 2024. <https://huggingface.co/datasets/wendlerc/RenderedText>.
- Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval, 2020. <https://arxiv.org/abs/2004.01804>.
- Luis Wiedmann, Aritra Roy Gosthipaty, and Andrés Marafioti. nanovlm. <https://github.com/huggingface/nanoVLM>, 2025.
- Qiyuan Wu, Chao Gao, Zhiqiang Fu, Wenxuan Li, Yang Yang, Yingqing Dong, Xiaoyi Wang, and Zhongyuan Zhang. Gui-actor: Coordinate-free visual grounding for gui agents. *arXiv preprint arXiv:2506.03143*, 2025. <https://arxiv.org/abs/2506.03143>.
- Siwei Wu, Kang Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, J. H. Liu, Ruibo Liu, Xingwei Qu, Xuxin Cheng, Ge Zhang, Wenhao Huang, and Chenghua Lin. MMRA: A benchmark for evaluating multi-granularity and multi-image relational association capabilities in large visual language models, 2024a. <http://arxiv.org/abs/2407.17379>.
- Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024b.
- Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding WordArt: Corner-guided transformer for scene text recognition. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13688, pages 303–321. Springer Nature Switzerland, 2022. ISBN 978-3-031-19814-4 978-3-031-19815-1. doi: 10.1007/978-3-031-19815-1_18. https://link.springer.com/10.1007/978-3-031-19815-1_18. Series Title: Lecture Notes in Computer Science.
- Anyi Xu and Qifeng Chen. Data selection for fine-tuning vision language models via cross-modal agreement and self-play. *arXiv preprint arXiv:2510.01454*, 2025. <https://arxiv.org/abs/2510.01454>.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions, 2025a. <https://arxiv.org/abs/2304.12244>.
- Peng Xu, Agrim Gupta, Cordelia Schmid, Marcus Rohrbach, and Wenhan Xiong. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv preprint arXiv:2212.10773*, 2023. <https://arxiv.org/abs/2212.10773>.

- Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguis: Unified pure vision agents for autonomous GUI interaction, 2025b. <http://arxiv.org/abs/2412.04454>.
- Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024. <http://arxiv.org/abs/2402.11690>.
- Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2051–2060, 2019. http://openaccess.thecvf.com/content_ICCV_2019/html/Yang_SpatialSense_An_Adversarially_Crowdsourced_Benchmark_for_Spatial_Relation_Recognition_ICCV_2019_paper.html.
- 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. <http://arxiv.org/abs/2502.14846>.
- Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model, 2023. <http://arxiv.org/abs/2310.05126>.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models, 2024a. <https://arxiv.org/abs/2309.12284>.
- Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, Yuning Du, Shikun Feng, Xiaoguang Hu, Pengyuan Lyu, Kun Yao, Yuechen Yu, Yuliang Liu, Wanxiang Che, Errui Ding, Cheng-Lin Liu, Jiebo Luo, Shuicheng Yan, Min Zhang, Dimosthenis Karatzas, Xing Sun, Jingdong Wang, and Xiang Bai. ICDAR 2023 competition on structured text extraction from visually-rich document images. In Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi, editors, *Document Analysis and Recognition - ICDAR 2023*, volume 14188, pages 536–552. Springer Nature Switzerland, 2023. ISBN 978-3-031-41678-1 978-3-031-41679-8. doi: 10.1007/978-3-031-41679-8_32. https://link.springer.com/10.1007/978-3-031-41679-8_32. Series Title: Lecture Notes in Computer Science.
- Youngjoon Yu, Sangyun Chung, Byung-Kwan Lee, and Yong Man Ro. SPARK: Multi-vision sensor perception and reasoning benchmark for large-scale vision-language models, 2024b. <http://arxiv.org/abs/2408.12114>.
- Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34(3):509–521, 2019. ISSN 1000-9000, 1860-4749. doi: 10.1007/s11390-019-1923-y. <http://link.springer.com/10.1007/s11390-019-1923-y>.
- Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4553–4562, 2022. http://openaccess.thecvf.com/content/CVPR2022/html/Yuan_Syntax-Aware_Network_for_Handwritten_Mathematical_Expression_Recognition_CVPR_2022_paper.html.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. MAMmoTH: Building math generalist models through hybrid instruction tuning, 2023. <http://arxiv.org/abs/2309.05653>.
- Bo-Wen Zhang, Yan Yan, Lin Li, and Guang Liu. Infinity-math: A scalable instruction tuning dataset in programmatic mathematical reasoning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5405–5409. ACM, 2024a. ISBN 979-8-4007-0436-9. doi: 10.1145/3627673.3679122. <https://dl.acm.org/doi/10.1145/3627673.3679122>.
- Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019. http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_RAVEN_A_Dataset_for_Relational_and_Analogical_Visual_REasoning_CVPR_2019_paper.html.
- Jingyang Zhang, Bo Li, Yanyuan Wang, Xuhong Wang, and Kai Chen. Self-filter: Instruction difficulty as a signal for multimodal fine-tuning. *arXiv preprint arXiv:2403.12776*, 2024b. <https://arxiv.org/abs/2403.12776>.
- Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024c.
- Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun

- Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. MAVIS: Mathematical visual instruction tuning with an automatic data engine, 2024d. <http://arxiv.org/abs/2407.08739>.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding, 2023. <http://arxiv.org/abs/2306.17107>.
- Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data, 2022. <http://arxiv.org/abs/2206.01347>.
- Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations, 2023. <http://arxiv.org/abs/2306.14321>.
- Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- Longtao Zheng, Zhiyuan Huang, Zhenghai Xue, Xinrun Wang, Bo An, and Shuicheng Yan. AgentStudio: A toolkit for building general virtual agents, 2024a. <http://arxiv.org/abs/2403.17918>.
- Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhui Chen, and Xiang Yue. OpenCodeInterpreter: Integrating code generation with execution and refinement, 2024b. <http://arxiv.org/abs/2402.14658>.
- Wu Zhiyong, He Zhenyu, Xu Fangzhi, Peng Chao, Jia Chengyou, Ding Zichen, Cui Junbo, Li Rui, Zhao Siheng, Xie Yuanming, You Wei, and Qiu Yao. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024. <https://arxiv.org/abs/2410.23218>.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance, 2021. <http://arxiv.org/abs/2105.07624>.
- Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866. ACM, 2022. ISBN 978-1-4503-9203-7. doi: 10.1145/3503161.3548422. <https://dl.acm.org/doi/10.1145/3503161.3548422>.
- Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023.
- Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. http://openaccess.thecvf.com/content_cvpr_2016/html/Zhu_Visual7W_Grounded_Question_CVPR_2016_paper.html.

A Appendix

A.1 Duplicate Cluster Visualization

Visualization of different results from the duplication detection pipeline. Choosing a single threshold to identify duplicated over multiple different categories is a balancing act between false-positives and false-negatives. After manual tuning we settled on $\tau = 0.95$.

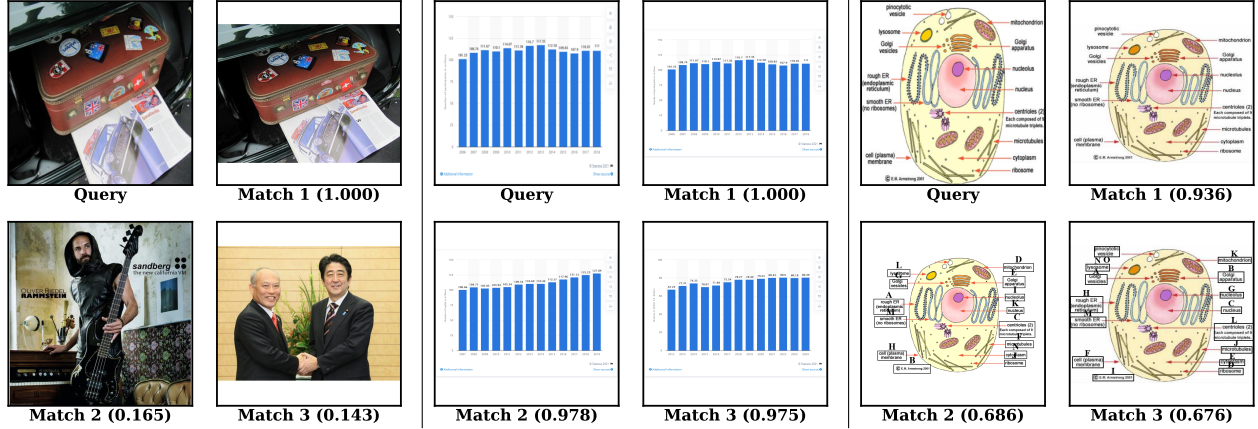


Figure 8 | Duplicate detection visualization with $\tau = 0.95$. Each panel shows the query image and retrieved matches with similarity scores. These three different scenarios, show the difficulty in picking a single threshold: (A, left) true photographic duplicates under mild crops/brightness; (B, middle) false positives (e.g., templated charts with different numbers) just above τ ; (C, right) false negatives (hand drawings) just below τ . After qualitative experiments we settled on $\tau = 0.95$ since it provided a good trade off between Precision and Recall.

A.2 Quality Ratings

These are the full prompts used with the LLM/VLM-as-a-judge pipeline to rate the quality of every turn in FineVision.

Relevance:

Rate how well this answer responds to the question (1-5):

- 5 - Excellent: Directly and completely answers the question with accurate, relevant information
- 4 - Good: Directly addresses the question with mostly relevant info, minor gaps acceptable
- 3 - Adequate: Partially addresses the question, some relevant information but incomplete
- 2 - Poor: Minimal attempt to answer, mostly irrelevant or significant gaps
- 1 - Inadequate: Completely unrelated, only meta-commentary, or unintelligible

RESPOND DIRECTLY WITH ONLY THE NUMBER. NO TEXT, NO EXPLANATION, JUST THE SCORE (1-5).

Question: {question}

Answer: {answer}

Score:

Formatting:

Rate the formatting quality of this text (1-5):

- 5 - Excellent: Clean, professional, proper grammar/punctuation, well-structured
- 4 - Good: Generally clean and readable, minor typos that don't impact understanding
- 3 - Acceptable: Readable despite some formatting issues, occasional special characters
- 2 - Poor: Significant formatting problems that impact readability, multiple errors
- 1 - Unacceptable: Severe corruption, extensive encoding errors, or completely garbled

RESPOND DIRECTLY WITH ONLY THE NUMBER. NO TEXT, NO EXPLANATION, JUST THE SCORE (1-5).

Question: {question}

Answer: {answer}

Score:

Visual Dependency:

Rate how much this question depends on visual information to be answered (1-5):

- 5 - Highly Visual: Requires specific visual details, asks about objects/scenes that must be seen
- 4 - Mostly Visual: Likely requires visual info, asks about visual properties or spatial relationships
- 3 - Moderately Visual: Could benefit from visual info but might be answerable with context
- 2 - Minimally Visual: Primarily answerable from general knowledge, visual info provides minor context
- 1 - Not Visual: Pure general knowledge, abstract concepts, no reference to visual elements

RESPOND DIRECTLY WITH ONLY THE NUMBER. NO TEXT, NO EXPLANATION, JUST THE SCORE (1-5).

Question: {question}

Score:

Image--Question Correspondence:

Rate how well this image corresponds to and supports answering the question (1-5):

- 5 - Perfect: Image directly contains all elements needed, ideal question-image pair
- 4 - Strong: Image contains most needed elements with clear visual information
- 3 - Moderate: Image contains some relevant info, partial match with reasonable connection
- 2 - Weak: Very limited relevant information, mostly unrelated content
- 1 - No Match: Completely unrelated, corrupted/blank image, or obvious mismatch

RESPOND DIRECTLY WITH ONLY THE NUMBER. NO TEXT, NO EXPLANATION, JUST THE SCORE (1-5).

Question: {question}

Score:

A.3 Action Space

Detailed description of the unified action space.

Category	Unified Actions
Shared (OS & Mobile)	<code>click(x: float, y: float)</code> <code>type(text: str)</code> <code>navigate_back()</code> <code>open_app(app_name: str)</code> <code>drag(</code> <code>from_coord: tuple[float, float],</code> <code>to_coord: tuple[float, float]</code> <code>)</code>
OS	<code>move_mouse(x: float, y: float)</code> <code>double_click(x: float, y: float)</code> <code>right_click(x: float, y: float)</code> <code>press(keys: str list[str])</code> <code>scroll(</code> <code>direction: Literal["up", "down", "left", "right"],</code> <code>amount: int</code> <code>)</code>
Mobile	<code>long_press(x: float, y: float)</code> <code>swipe(</code> <code>from_coord: tuple[float, float],</code> <code>to_coord: tuple[float, float]</code> <code>)</code>
Completion	<code>final_answer(answer: str)</code> <code>wait(seconds: int)</code>

Table 3 | Unified action space schema with categories and typed arguments.

A.4 Data Quality Filtering.

We evaluated our simple prompt-based quality scores as filters along the four axes defined in Sec. 3.2. Across our experiments, these specific scores did not yield an effective filtering scheme: reducing the dataset by thresholding on these metrics generally did not improve model performance compared to training on the unfiltered data (see Fig. 9 and 10). This stands in contrast to recent works that report measurable gains from explicit multimodal data selection/cleaning: *Eagle2* (Li et al., 2025b) applies rule-based filtering and mixture shaping over large pools; *XMAS* (Xu and Chen, 2025) selects via cross-modal agreement and self-play; and *Self-Filter* (Zhang et al., 2024b) retains the most challenging instructions via difficulty-aware selection. Our negative result therefore suggests that how the filter is constructed matters: naive prompt-score thresholding (as instantiated here) is insufficient, whereas targeted procedures (e.g., consistency/difficulty scoring, concept balancing, deduplication) can be beneficial. Moreover, the goal of filtering itself warrants scrutiny, as some common practices can be actively harmful. For instance, Pouget et al. (2024) demonstrate that filtering web-scale data to English-only pairs degrades a model’s cultural understanding and harms performance for underrepresented socioeconomic groups, even while boosting scores on Western-centric benchmarks. We therefore conclude only that our prompt-based quality metrics, as instantiated here, are not good filters; we do not make claims about other quality estimators or alternative filtering strategies. To facilitate further work, we release per-turn scores so others can explore different uses or models for data selection.

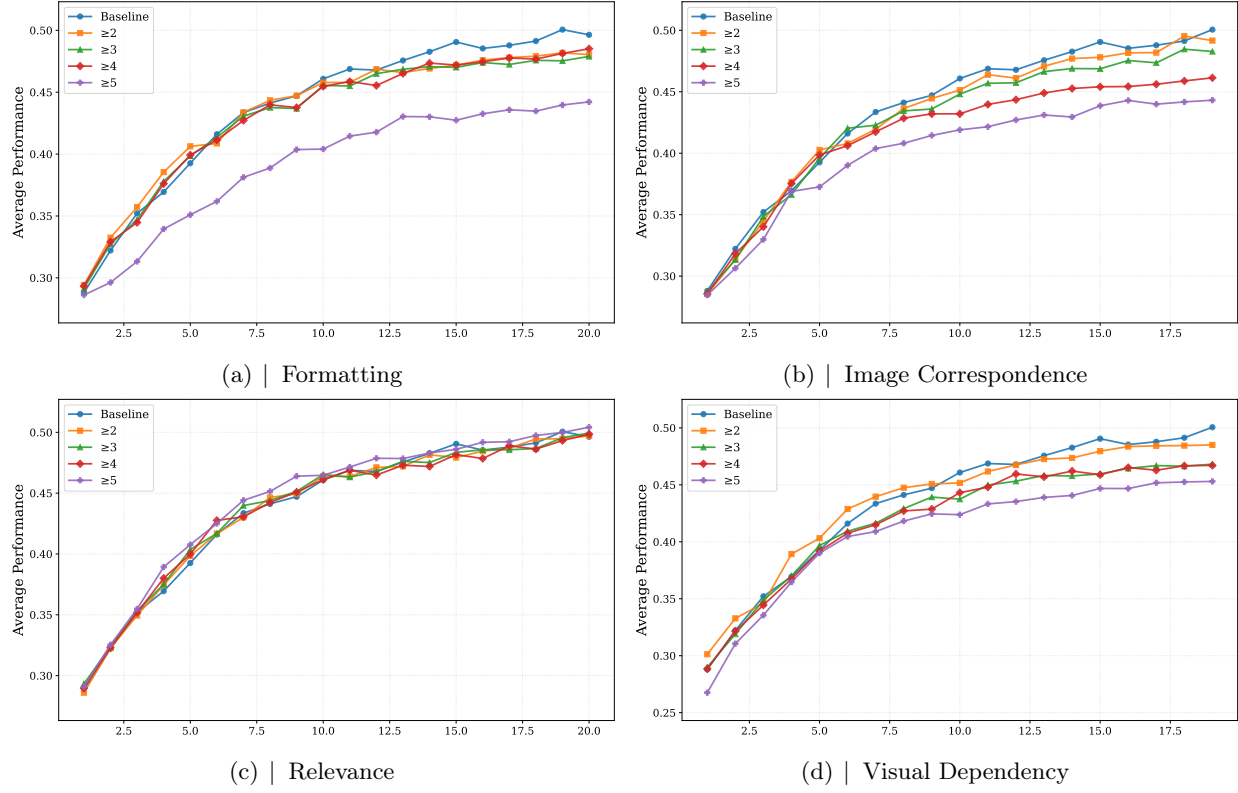


Figure 9 | Model performance under prompt-based quality filtering. Average benchmark performance for models trained with thresholds on our four prompt-based quality axes. These results indicate that our specific prompt-based scores are not effective data filters.

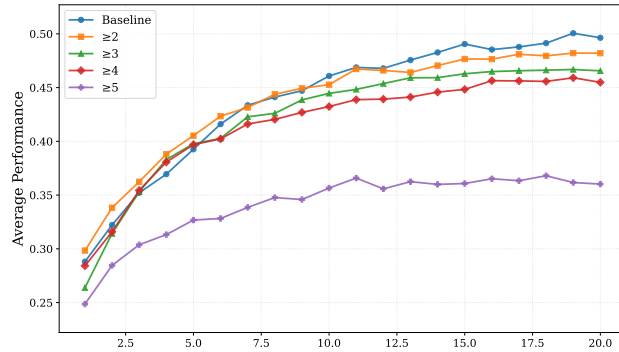


Figure 10 | Model performance under combined prompt-based quality filtering. We we combine all filters into a single criterion, meaning we only select datapoints that have all four ratings above a certain threshold, training on the full dataset also results in the best performance.

A.5 Benchmark Contamination and Effect

Detailed statistics regarding the benchmark contamination as well as the performance drop after removing these samples.

Name	Samples	Contamination Rate	Performance Drop
Cauldron	1.8M	3.05%	2.8%
LLaVA-Vision	3.9M	2.15%	2.7%
Cambrian-7M	7.0M	2.29%	3.7%
FineVision	24.3M	1.02%	1.6%

Table 4 | Contamination and performance drop across datasets.

A.6 Additional Statistics: Token Length, Conversation Turns and Image Resolution by Category

The split-violin plots in Fig. 11 show how interaction type shapes sequence length. Questions are short and tightly concentrated across categories, whereas answers are broader and often heavy-tailed. These shapes yield three archetypes: *perceptual/extractive* (**Grounding**, **General VQA**, **Chart & Table**) with compact distributions; *descriptive generation* (**Captioning**) with no question and medium-length captions; and *transcription* (**Naive OCR**), long tail but driven by fidelity rather than inference. We include both **Naive OCR** (e.g., “What is the text in the image?”) and **OCR QA** (questions that require reading to be answered), treating the latter as more involved reading comprehension and for whole documents.

The disparity between short prompts and longer answers is an information gap the model must fill. It is largest for **Naive OCR/OCR QA**, moderate for **Science**, and minimal for **Grounding** and **Chart & Table**. Typical median answer-minus-question token gaps by category (dotted lines in Fig. 11) are: Text-only 85.99, Science 33.51, OCR QA 15.33, Naive OCR 104.15, Mathematics 1.36, Grounding & Counting 12.85, General VQA 45.64, Chart & Table −19.26, and Captioning & Knowledge 203.47. **Chart & Table** - and, to a lesser extent, **Grounding & Counting / OCR QA** - naturally support multi-turn exchanges because several queries can target the same figure/document/screenshot. See also the distribution of turns per sample by category in Fig. 12. Image resolutions are broadly similar across categories, with document-centric categories (e.g. **OCR QA**) skewing higher to preserve legibility. Medians are post-resizing and pictured as dotted lines in the figures; median (width, height) by category are: Science (485.03, 332.58), OCR QA (1189.77, 1428.63), Naive OCR (700.56, 443.41), Mathematics (755.31, 608.22), Grounding (1642.53, 950.31), General VQA (641.29, 515.80), Chart & Table (832.69, 600.97), and Captioning (796.18, 629.51).

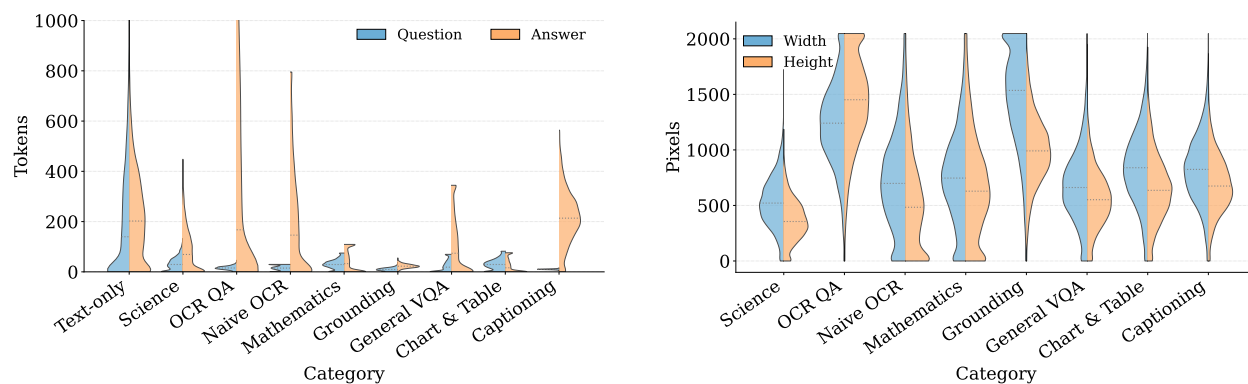


Figure 11 | Token length and image resolution by category. Left: split-violin token-length distributions by category; questions (blue) vs. answers (orange), median is dotted line, y-axis capped at 1000 for visibility. These shapes expose task archetypes and the information gap between prompt and response. Right: image resolution distributions by category; width (blue) and height (orange), median is dotted line.

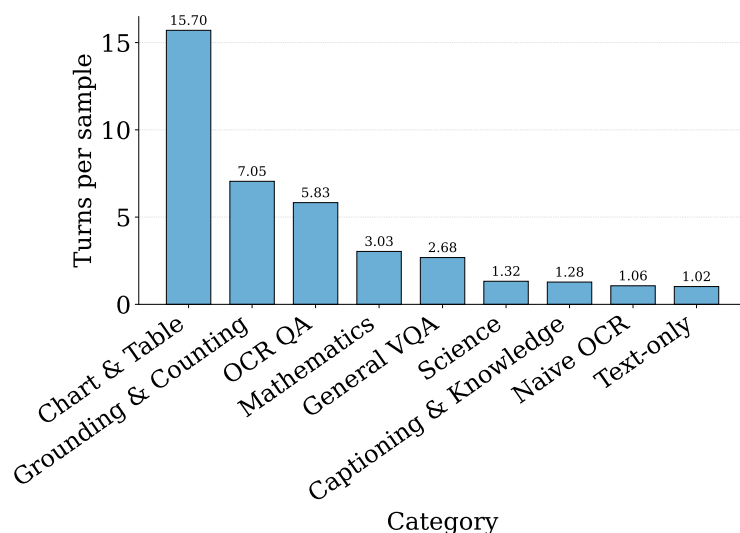


Figure 12 | Turns per sample by category. Categories such as Chart & Table and Grounding & Counting support more multi-turn interactions per image.

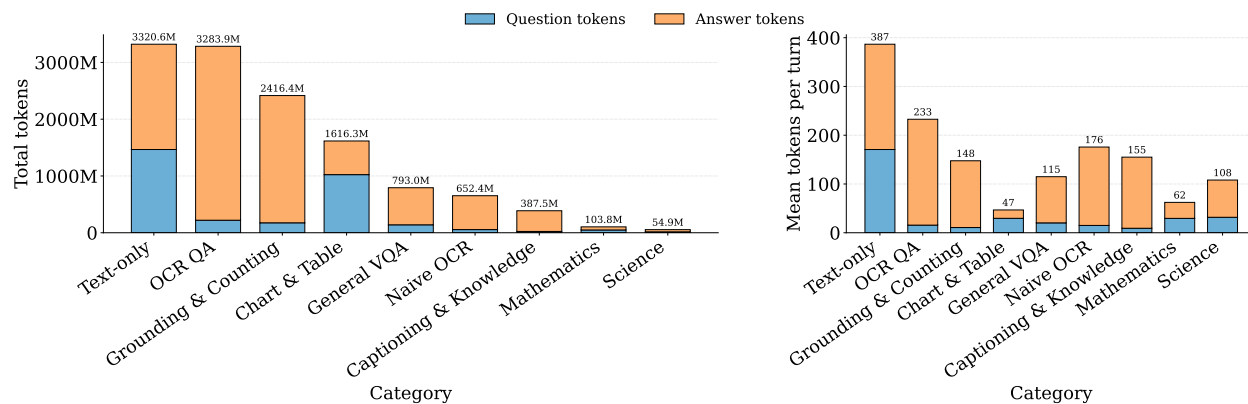


Figure 13 | Total tokens by category stacked by question and answer. Left: total tokens, Right: mean tokens per turn

A.7 FineVision Dataset Subsets

Detailed description and statistics of the FineVision dataset subsets by category (see Tables 5, 6, 7, 8, 9, 10, 11, 12, 13).

Subset Name	Images	Samples	Turns	Answer Tokens
coco_colors (Karakus, 2024)	118287	118287	118287	6376672
densefusion_1m (Li et al., 2024b)	1058751	1058751	1058751	263718217
face_emotion (FastJobs, 2024)	797	797	797	8066
google_landmarks (Weyand et al., 2020)	299993	299993	842127	10202980
image_textualization(filtered) (Pi et al., 2024)	99573	99573	99573	19374090
laion_gpt4v (LAION, 2023)	9301	9301	9301	1875283
localized_narratives (Pont-Tuset et al., 2020)	199998	199998	199998	8021473
sharegpt4o (OpenGVLab, 2024)	57284	57284	57284	36555323
sharegpt4v(coco) (Chen et al., 2024b)	50017	50017	50017	9825387
sharegpt4v(knowledge) (Chen et al., 2024b)	1988	1988	1988	293850
sharegpt4v(llava) (Chen et al., 2024b)	29986	29986	29986	6175899
sharegpt4v(sam) (Chen et al., 2024b)	8990	8990	8990	1668797
textcaps (Sidorov et al., 2020)	21906	21906	21906	355991

Table 5 | Captioning & Knowledge datasets

Subset Name	Images	Samples	Turns	Answer Tokens
aguvis-stage-1 (Xu et al., 2025b)	458957	458957	3831666	93546182
groundui (Zheng et al., 2024a)	13531	13531	18016	883274
objects365_qa (Shao et al., 2019)	1742287	1742287	12329259	2146619635
oodvqa (Tu et al., 2023)	8488	8488	8488	8488
tallyqa (Acharya et al., 2019)	98680	98680	183986	370282

Table 6 | Grounding & Counting datasets

Subset Name	Images	Samples	Turns	Answer Tokens
ai2d_merged (Kembhavi et al., 2016)	4858	4858	12325	1319140
CoSyn_400k_chemical (Yang et al., 2025)	8942	8942	55391	2450290
CoSyn_400k_circuit (Yang et al., 2025)	10470	10470	67939	2637618
pathvqa (He et al., 2020)	32632	32632	32632	85168
pmc_vqa(mathv360k) (Shi et al., 2024)	35948	35948	35948	255109
scienceqa (Lu et al., 2022)	4976	4976	6149	18447
scienceqa(nona_context) (LMMS-Lab, 2025)	19208	19208	19208	25311
tqa (Kembhavi et al., 2017)	2749	2749	12567	149776
visualwebinstruct(filtered) (Jia et al., 2025)	263581	263581	263581	31802459
vqarad (Lau et al., 2018)	313	313	1793	6003

Table 7 | Science datasets

Subset Name	Images	Samples	Turns	Answer Tokens
geoqa+(mathv360k) (Cao and Xiao, 2022)	17162	17162	17162	117740
unigeo(mathv360k) (Chen et al., 2022a)	11949	11949	11949	81781
clevr (Lindström and Abraham, 2022)	70000	70000	699989	1570525
clevr_math (Lindström and Abraham, 2022)	70000	70000	556082	580324
clevr_math(mathv360k) (Shi et al., 2024)	5280	5280	5280	27536
CoSyn_400k_math (Yang et al., 2025)	66714	66714	66714	28631388
geo170k(aligned) (Gao et al., 2023)	35297	35297	35297	1866019
geo170k(qa) (Gao et al., 2023)	12101	12101	12101	1115242
geo3k (Lu et al., 2021a)	2091	2091	2091	2091
geometry3k(mathv360k) (Shi et al., 2024)	9724	9724	9724	69075
geomverse (Kazemi et al., 2023)	9303	9303	9339	2454014
geos(mathv360k) (Seo et al., 2015)	498	498	498	3509
intergps (Lu et al., 2021a)	1280	1280	1760	5280
mavis_math_metagen (Zhang et al., 2024d)	87348	87348	87348	5486485
mavis_math_rule_geo (Zhang et al., 2024d)	99986	99986	99986	12535251
raven (Zhang et al., 2019)	63081	42000	42000	63081
super_clevr(mathv360k) (Li et al., 2023)	8642	8642	8642	44129

Table 8 | Mathematics datasets

Subset Name	Images	Samples	Turns	Answer Tokens
text_ruozhiba (LooksJuicy, 2024)	0	1496	1496	234822
text_code_feedback (Zheng et al., 2024b)	0	66383	221096	79752351
text_codefeedback_filtered_instruction (Zheng et al., 2024b)	0	156525	156525	62764414
text_infinitymath (Zhang et al., 2024a)	0	101380	101380	212543
text_mathinstruct (Yue et al., 2023)	0	262039	262039	44145362
text_mathqa (Yu et al., 2024a)	0	394996	394996	72451061
text_mathstepdpo10k (Lai et al., 2024)	0	10795	10795	989312
text_numinamath_cot (LI et al., 2024)	0	859494	859494	387758581
text_openhermes_2_5 (Teknium, 2023)	0	1001551	1008268	233561291
text_openorca (Lian et al., 2023)	0	4233853	4233853	468042176
text_orcamath (Mitra et al., 2024)	0	200035	200035	61860987
text_pythoncode25k (FLOCK4H, 2024)	0	49626	49626	4945892
text_pythoncodealpaca (Bisht, 2024)	0	18612	18612	2683469
text_theoremqa (Chen et al., 2023)	0	800	800	3468
text_wizardlm_evol (Xu et al., 2025a)	0	69999	69999	21955856
text_OpenMathInstruct-2 (Toshniwal et al., 2024)	0	1000000	1000000	413132418

Table 9 | Text-only datasets

Subset Name	Images	Samples	Turns	Answer Tokens
Unichart (Masry et al., 2023)	611925	611925	6898324	211989247
tat_dqa (Zhu et al., 2022)	2448	2207	13251	1177852
chart2text (Kantharaj et al., 2022)	26961	26961	30215	2670580
chartqa (Masry et al., 2022)	18265	18265	28287	134793
CoSyn_400k_chart (Yang et al., 2025)	116814	116814	1085882	57641030
CoSyn_400k_table (Yang et al., 2025)	46518	46518	416519	23335054
dvqa (Kafle et al., 2018)	200000	200000	2325316	5477966
figureqa (Kahou et al., 2017)	100000	100000	1327368	2654736
figureqa(mathv360k) (Shi et al., 2024)	17587	17587	17587	97404
finqa (Chen et al., 2022b)	5276	5276	6251	224015
hitab (Cheng et al., 2022)	2500	2500	7782	335013
lrn_chart (Li et al., 2024)	1776	1776	5372	158711
mmc_instruct (Liu et al., 2023c)	168178	168178	168178	74581055
multihiertr (Zhao et al., 2022)	30875	7619	7830	244744
plotqa (Methani et al., 2020)	157070	157070	20249479	118122387
robut_sqa (Zhao et al., 2023)	8514	8514	34141	1794570
robut_wikisql (Zhao et al., 2023)	74989	74989	86202	9276100
robut_wtq (Zhao et al., 2023)	38246	38246	44096	6415830
SynthChartNet (Nassar et al., 2025)	500000	500000	500000	67392223
tabmwp (Lu et al., 2023)	22722	22722	23021	1883243
tabmwp(mathv360k) (Shi et al., 2024)	22452	22452	22452	158042
tat_qa (Zhu et al., 2021)	2199	2199	13215	254790
vistext (Tang et al., 2023)	9969	9969	9969	1191127
vqaonbd (VQAonBD, 2023)	39986	39986	1254165	5620523

Table 10 | Chart & Table datasets

Subset Name	Images	Samples	Turns	Answer Tokens
alfworldgpt (Shridhar et al., 2021)	45073	45073	45073	6276573
chinesememe (Contributors, 2024)	54212	54212	54212	21122723
wildvision (Lu et al., 2024)	333	333	405	72820
allava_laion (Chen et al., 2024a)	468664	468664	937328	145799426
allava_vflan (Chen et al., 2024a)	177078	177078	387872	55305642
LLaVA_Instruct_150K (Liu et al., 2023d)	157710	157710	361405	28719278
datik (Belouadi et al., 2023)	220537	222385	222385	187757952
cambrian(filtered)_processed (Tong et al., 2024)	83123	83124	98534	5503211
cocoqa (Ren et al., 2015)	46287	46287	78736	212480
CoSyn_400k_graphic (Yang et al., 2025)	26968	26968	26968	8235679
datikz (Belouadi et al., 2023)	47441	47974	48296	59116193
drivelm (Sima et al., 2024)	90049	4072	161030	1431417
hateful_memes (Kiela et al., 2020)	8500	8500	8500	17000
iconqa (Lu et al., 2021b)	27307	27307	29841	72492
iconqa(mathv360k) (Shi et al., 2024)	22589	22589	22589	134029
idk (Cha et al., 2024)	11123	11123	27614	665247
indoor_qa (Kerem, 2024)	3350	3350	3350	19700
llavar_gpt4_20k (Zhang et al., 2023)	19790	19790	43167	1516730
lnqa (Pont-Tuset et al., 2020)	302780	302780	1520942	19027663
lrv_normal(filtered) (Liu et al., 2023b)	10489	10489	155269	3134247
lvis_instruct4v (Wang et al., 2023b)	222711	222711	1050622	43726782
mimic_cgd (Laurençon et al., 2024)	141878	70939	141869	4304380
mmevol (Luo et al., 2024)	160215	160215	630441	50445237
mmra (Wu et al., 2024a)	2048	1024	1024	25764
nlvr2 (Suhr et al., 2017)	100852	50426	86373	172746
sketchyvqa (Tu et al., 2023)	8000	8000	8000	8000
spark (Yu et al., 2024b)	3904	3904	6248	73973
spatialsense (Yang et al., 2019)	10440	10440	17498	418883
spot_the_diff (Jhamtani and Berg-Kirkpatrick, 2018)	17132	8566	9524	209630
vision_flan(filtered) (Xu et al., 2024)	175964	175964	175964	3009891
visual7w (Zhu et al., 2016)	14366	14366	69817	209451
vizwiz(mathv360k) (Gurari et al., 2018)	6604	6604	6604	44876
vqav2 (Goyal et al., 2017)	82772	82772	443757	1100837
vsr (Liu et al., 2023a)	2157	2157	3354	6708
websight (Laurençon et al., 2024)	10000	10000	10000	5237381
yesbut (Nandy et al., 2024)	4318	4318	4318	157229

Table 11 | General VQA datasets

Subset Name	Images	Samples	Turns	Answer Tokens
ctw (Yuan et al., 2019)	24290	24290	180621	1653254
k12_printing (TAL, 2023)	256636	256636	256636	7465001
svrd (Yu et al., 2023)	4396	4396	4396	834514
tal_ocr_eng (TAL, 2023)	256646	256646	256646	7465207
mathwriting-google (Gervais et al., 2025)	300000	300000	300000	5954806
art (Chng et al., 2019)	5603	5603	5603	283138
captcha (Samadnejad, 2024)	113062	113062	113062	466856
chrome_writting (Mouchere et al., 2013)	8825	8825	8825	172940
cocotext (Veit et al., 2016)	16169	16169	16169	177111
funsd (Jaume et al., 2019)	194	194	3879	29996
hme100k (Yuan et al., 2022)	74492	74492	74492	1757743
hw_squad (Mathew et al., 2021a)	20457	20457	83682	388518
iam (Marti and Bunke, 2002)	5663	5663	5663	130794
iiit5k (Mishra et al., 2012)	1990	1990	1990	4259
imgur5k (Krishnan et al., 2023)	5934	5934	5934	288054
latex_handwritten (Mouchere et al., 2013)	39583	39583	39583	1874733
latexformulas (OleehyO, 2024)	552340	552340	552340	43094747
maptext (Li et al., 2024c)	200	200	799	70813
memotion (Ramamoorthy et al., 2022)	6991	6991	6991	177429
orand_car_a (Diem et al., 2014)	1999	1999	1999	9035
rendered_text (Wendler, 2024)	10000	10000	10000	244183
sroie (Huang et al., 2019)	33616	33616	33616	243240
SynthCodeNet (Nassar et al., 2025)	499983	499983	499983	253422136
synthdog (Kim et al., 2022)	500000	500000	500000	48010145
SynthFormulaNet (Nassar et al., 2025)	499997	499997	499997	51215097
wordart (Xie et al., 2022)	19066	4804	4804	54263
olmOCR-mix-0225-documents (Poznanski et al., 2025)	228864	228864	228858	163194337
olmOCR-mix-0225-books (Poznanski et al., 2025)	15194	15194	15194	7962779

Table 12 | Naive OCR datasets

Subset Name	Images	Samples	Turns	Answer Tokens
a_okvqa (Schwenk et al., 2022)	54602	54602	54602	360990
est_vqa (Wang et al., 2020)	19358	19358	19358	143270
mmsoc_memotion (Ramamoorthy et al., 2022)	6991	6991	6991	421250
arxivqa (Li et al., 2024a)	100000	100000	100000	6422269
DoclingMatix (Nassar et al., 2025)	2465202	1270911	10626898	2996338775
ureader_qa_processed (Ye et al., 2023)	252953	252953	252953	930617
aokvqa (Schwenk et al., 2022)	16539	16539	17056	218917
bentham (Mathew et al., 2021a)	10843	10843	10843	124459
blockdiagramcomputerized (Bhushan and Lee, 2022)	502	502	502	34453
blockdiagramhandwritten (Bhushan and Lee, 2022)	1029	1029	1029	75598
CoSyn_400k_diagram (Yang et al., 2025)	34963	34963	300357	11943321
CoSyn_400k_document (Yang et al., 2025)	71282	71282	605173	16095526
CoSyn_400k_music (Yang et al., 2025)	11969	11969	81786	3175586
CoSyn_400k_nutrition (Yang et al., 2025)	6931	6931	112097	3687254
diagram_image_to_text (Kamizuru, 2024)	300	300	300	20723
docvqa (Mathew et al., 2021b)	10189	10189	39463	275510
handwriting_forms (Forms, 2024)	1400	1400	1400	41490
infographic_vqa (Mathew et al., 2022)	1982	4394	23717	86951
infographic_vqa_llava_format (Mathew et al., 2022)	4394	2113	10054	43912
infographic(gpt4v) (Mathew et al., 2022)	2113	1982	1982	1044183
invoices_receipts (Minyang, 2024)	3013	3013	3013	771948
mapqa (Chang et al., 2022)	37417	37417	483416	5657339
mapqa(mathv360k) (Shi et al., 2024)	5225	5225	5225	44560
ocrvqa (Mishra et al., 2019)	165746	165746	801579	4801833
pdfvqa (Ding et al., 2023)	8593	8593	95000	939948
screen2words (Wang et al., 2021)	15730	15730	15743	120781
screenqa (Hsiao et al., 2022)	80761	80761	80761	826795
slidevqa (Tanaka et al., 2023)	11868	1919	10617	156036
st_vqa (Biten et al., 2019)	17247	17247	23121	98892
sujet_finance (Sujet AI, 2024)	9801	9801	107050	1925361
textocr(gpt4v) (Carter, 2024)	25060	25060	25060	2436974
textvqa (Singh et al., 2019)	21953	21953	34602	141882
ureader_cap (Ye et al., 2023)	91215	91215	91215	1435964
ureader_ie (Ye et al., 2023)	17320	17320	17320	128229
ureader_kg_processed (Ye et al., 2023)	37550	37550	37550	2013731
visualmrc (Tanaka et al., 2021)	3027	3027	11988	147385

Table 13 | OCR QA datasets