20/20 Vision Language Models:

A Prescription for Better VLMs through Data Curation Alone

DatologyAI Team[1]

Data curation has shifted the quality–compute frontier for language-model and contrastive image-text pretraining, but its role for vision-language models (VLMs) is far less established. Here, we ask how far data curation alone can take VLM performance, holding architecture, training recipe, and compute fixed and varying only the training data. Our pipeline, applied to the MAmmoTH-VL single-image subset, lifts performance by +11.7**pp** on average across 20 public VLM benchmarks (spanning grounding, VQA, OCR and documents, captioning, spatial and 3D, counting, charts, math, brand-ID, and multi-image reasoning) and by +11.3**pp** on average across all nine capability axes of **DatBench** (Joshi et al., 2026), our high-fidelity, comprehensive VLM eval suite. At 2B, our *curated* model surpasses InternVL3.5-2B by 9.9pp at $\sim 17\times$ less training compute and closes the gap to Qwen3-VL-2B to within 1.8pp at $\sim 87\times$ less compute, all from pretraining alone. Beyond strong gains in accuracy, our curation pipeline delivers four further properties: (1) **Reliability**: per-capability standard deviation across training seeds drops by $\sim 67\%$ and the curated-vs-baseline lift survives a context-length sweep from 4k to 16k tokens; (2) **OOD generalization**: the 9-eval OOD average rises by +7.2pp, and multi-image reasoning on BLINK rises by +3.09pp despite single-image-only training, with Visual Correspondence gaining +11.8pp despite demanding cross-image reasoning; (3) **Behavioral gains beyond benchmarks**: across ∼1,100 open-ended queries the *curated* 2B is more honest and more specific than the matched-compute *baseline*, and answers more concisely and refuses fewer benign queries than a frontier 2B reference; (4) **Pareto-dominance on inference cost**: at every scale tested (1B, 2B, 4B) the *curated* model raises accuracy while lowering response FLOPs against the matched-compute *baseline*, and the *curated* 4B matches near-frontier accuracy at $3.3\times$ lower response FLOPs than Qwen3-VL-4B. Our pipeline shows that data curation is a high-leverage tool for building better VLMs, reaching near-frontier accuracy at up to $\sim 150\times$ less training compute.

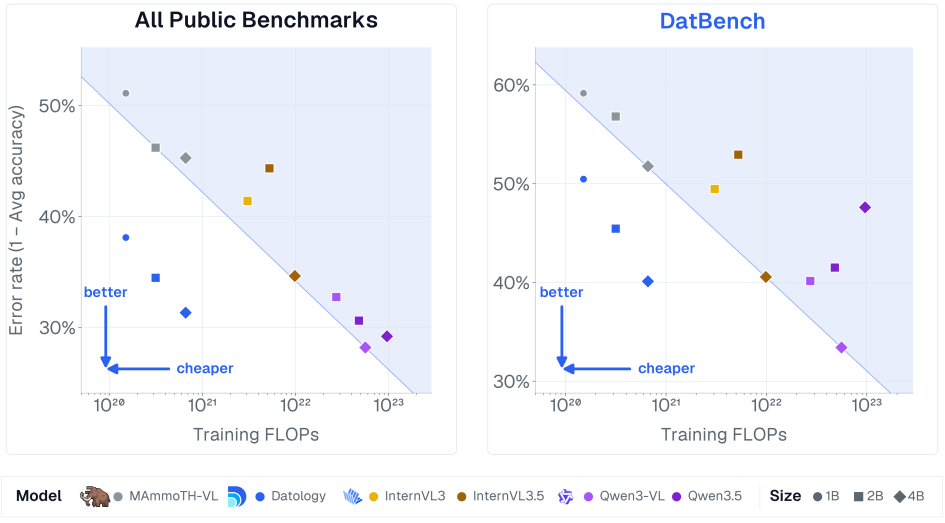


Figure 1 Curation matches frontier 2B/4B VLMs at a fraction of the training compute. Pareto frontier of error rate (1− avg. accuracy) vs. training FLOPs on (left) the 20-eval public VLM benchmark suite and (right) **DatBench**. DatologyAI-curated models (blue) at 1B, 2B, and 4B Pareto-dominate the matched-compute MAmmoTH-VL *baseline* (gray) at every scale, and approach Qwen3-VL, InternVL3.5, and Qwen3.5 (extensively post-trained, up to $\sim 150\times$ more compute) on both suites.

[1]See Contributions and Acknowledgements (§ 9) for full author list.

1 Introduction

In the era of deep learning, curating training data has been a key lever for improvements across modalities. The efficacy of training data curation has been extensively studied in language-model pretraining: curated open corpora (DCLM, FineWeb, RedPajama, Dolma, Nemotron-CC) can shift the quality–compute Pareto frontier by as much as architectural or scale choices (Li et al., 2024a; Penedo et al., 2024; Weber et al., 2024; Soldaini et al., 2024; Su et al., 2024a; DatologyAI, 2024b; Maini and DatologyAI Team, 2025). Data filtering and mixture design have also become first-class axes in contrastive image-text training (Gadre et al., 2023; Xu et al., 2023; Fang et al., 2024; DatologyAI, 2024a). In vision-language model research, by contrast, the training mixture is typically characterized only at the level of "X billion image–text pairs" rather than treated as a primary design variable, with most attention paid to non-data variables such as encoder choice, connector design, resolution handling, visual instruction tuning, and reinforcement learning (Liu et al., 2023a; Chen et al., 2024d; Wang et al., 2024a; Tong et al., 2024a; McKinzie et al., 2024; Deitke et al., 2024; Yu et al., 2024; Shen et al., 2025; Liu et al., 2025b). Holding architecture, training recipe, and compute fixed, we measure the impact of pretraining data curation alone on VLM capability.

Prior work on VLM data curation has typically isolated a single axis at a time: multimodal deduplication (Slyman et al., 2024; Liu et al., 2025a), image-text quality filtering (Wang et al., 2024b, 2025b; Chen et al., 2024b), data-mixture design (McKinzie et al., 2024; Tong et al., 2024a; Laurençon et al., 2024b; Deitke et al., 2024; Shi et al., 2024; Chen et al., 2024c; Wang et al., 2025a), recaptioning of paired data (Chen et al., 2023; Li et al., 2024b), task-agnostic synthetic data (Liu et al., 2024c; Su et al., 2024b; Liao et al., 2025), or task-specific synthetic data (Joshi et al., 2025b; Yang et al., 2025). Others release curated corpora without a matched-compute uncurated baseline (Guo et al., 2024). We instead compose joint image-and-text deduplication, joint image-and-text filtering, target-distribution matching for mixture design, and both task-agnostic and task-specific synthetic data into a single end-to-end pipeline, and ask how far data curation alone can take VLMs.

Concretely, we run our end-to-end data curation pipeline on the single-image subset (10M) of MAmmoTH-VL-12M (Guo et al., 2024) and train models across 1B–4B language backbones (Qwen3 LM (Qwen Team, 2025a) with SigLIP2 vision encoder (Tschannen et al., 2025)). We compare against a *baseline* trained on the same uncurated MAmmoTH-VL subset under identical training recipe and compute, and against SOTA public models at similar scales: the Qwen3-VL family (2B/4B) (Qwen Team, 2025b), the InternVL3 and InternVL3.5 families (2B/4B) (Chen et al., 2024d), and Qwen3.5 (2B/4B) (Qwen Team, 2026). These are strong reference points: trained on up to $\sim$150$\times$ more compute and including extensive post-training (instruction tuning, RLHF, RLVR) on top of their pretraining stage.

At 2B, our curation lifts the average across 20 public VLM benchmarks by +11.7**pp**, spanning grounding, VQA, OCR and documents, captioning, spatial and 3D reasoning, counting, charts, math, brand-ID, and multi-image reasoning, and the average across all nine capability axes of **DatBench** by +11.3**pp**. Figure 1 previews the cost-quality Pareto: the *curated* models Pareto-dominate the matched-compute *baseline* at every scale tested (1B, 2B, 4B), and approach extensively post-trained references at up to $\sim$ 150$\times$ less training compute. All results below are post-decontamination, applied to *baseline* and *curated* runs alike; the multimodal decontamination methodology is detailed in Appendix A.

We organize the paper around five key findings on the impact of data curation for VLMs:

- **Curation shifts the cost-quality Pareto frontier across model scales.** (§4) Across the 20-eval public suite, curation improves accuracy at every tested scale: +13.0**pp** at 1B, +11.7**pp** at 2B, and +14.0**pp** at 4B. Across **DatBench**'s nine capability axes, the gains are similarly large, reaching +11.3**pp** at 2B and increasing with scale. Moreover, these gains close the cost-quality gap between pretraining-only curation and extensively post-trained references.

- **Strong OOD generalization.** (§4) Our curation lifts the 9-eval OOD average by +7.2**pp** at 2B. Gains extend even to capabilities absent from training: multi-image reasoning on BLINK rises by +3.09**pp** overall, with `Visual_Correspondence` gaining +11.8**pp** despite demanding cross-image reasoning.
- **Reliable across seeds and robust across context lengths.** (§5) On top of the mean lift, per-capability standard deviation across training seeds drops from 2.47 to 0.82pp ($\sim$ 67% reduction), and the advantage survives a context-length sweep from 4k to 16k tokens.
- **Gains carry through to open-ended use beyond benchmarks.** (§6) On ∼1,100 open-ended queries spanning OCR, brand and franchise recognition, scene description, and refusal calibration, the *curated* 2B is more honest and more specific than the matched-compute *baseline*, and answers more concisely and refuses fewer benign queries than a frontier 2B reference.
- **Pareto-dominant on inference cost.** (§7) *Curated* models simultaneously raise accuracy and lower response-FLOPs cost against the matched-compute *baseline* at every scale tested (1B, 2B, 4B). Against extensively post-trained references, the *curated* 4B reaches 68.7% average accuracy at 3.3× lower response FLOPs than Qwen3-VL-4B, and exceeds Qwen3-VL-2B at 1.5× lower response FLOPs.

Together, these results show that pretraining data curation alone produces VLM gains that are (1) large, (2) OOD-generalizing, (3) reliable, (4) evident beyond benchmarks, and (5) inference-efficient, at up to $\sim$ 150× less training compute than extensively post-trained references.

2 Related Work

We organize related work along the three axes most relevant to this work. *VLM modeling* covers the dominant non-data levers vision-language modeling research has emphasized. *VLM evaluation* surveys the benchmark landscape. *Data curation* situates our contribution within the broader curation literature, spanning classical coreset methods, LM and contrastive pretraining, and VLM-specific work.

VLM modeling. A now-standard recipe couples a pretrained vision encoder with a pretrained language-model backbone via a lightweight projector, trained on interleaved or paired image-text data (Liu et al., 2023b,a; Li et al., 2023b; Alayrac et al., 2022; Dai et al., 2023). Architectural variants differ primarily in how visual tokens are produced and consumed: resampler-style cross-attention (Alayrac et al., 2022; Laurençon et al., 2024b), linear or MLP projectors over patch tokens (Liu et al., 2023a), dynamic or any-resolution tiling (Liu et al., 2024a; Chen et al., 2024d; Bai et al., 2023; Wang et al., 2024a; Qwen Team, 2025b), and mixture-of-encoder designs (Tong et al., 2024a; Shi et al., 2024; McKinzie et al., 2024). Recent open-weight families in the 2B–8B class (the Qwen-VL line, InternVL, Idefics, Molmo, MM1, PaliGemma, Cambrian, NVLM) (Wang et al., 2024a; Qwen Team, 2025b; Chen et al., 2024d,c; Laurençon et al., 2024b,a; Deitke et al., 2024; McKinzie et al., 2024; Beyer et al., 2024; Tong et al., 2024a; Dai et al., 2024) center most of their work on encoder choice, connector design, resolution handling, and post-training, with a parallel line targeting capabilities downstream of pretraining via preference optimization and reinforcement learning across hallucination, visual reasoning, grounding, counting, and spatial tasks (Yu et al., 2024; Sun et al., 2023; Zhou et al., 2024; Li et al., 2023c; Huang et al., 2025; Shen et al., 2025; Meng et al., 2025; Liu et al., 2025b). Several of the open-weight families do include mixture or data-inclusion ablations as pre-work (McKinzie et al., 2024; Tong et al., 2024a; Deitke et al., 2024; Laurençon et al., 2024b), but these stay inside a single model recipe and reduce the data question to whether one mixture improves an aggregate benchmark score. Our focus is the complementary question of isolating pretraining data as a design variable while holding architecture, recipe, and compute fixed.

VLM evaluation. Existing benchmarks cover knowledge and reasoning (Yue et al., 2024; Chen et al., 2024a; Liu et al., 2023c; Fu et al., 2023; Li et al., 2023a; Yu et al., 2023), perception and multi-image reasoning (Fu et al., 2024; Tong et al., 2024a; xAI, 2024), optical character recognition (OCR) and document understanding (Liu et al., 2024b; Mathew et al., 2021), hallucination (Li et al., 2023d), counting (Paiss et al., 2023; Acharya

et al., 2019), and grounding (Kazemzadeh et al., 2014; Yu et al., 2016). Parallel work has documented saturation, prompt sensitivity, and contamination in these suites (Chen et al., 2024a; Tong et al., 2024b; Adiga et al., 2025), motivating more targeted diagnostic benchmarks. **DatBench** (Joshi et al., 2026) builds on this line and is the evaluation instrument we use throughout the paper.

Data curation. In contrastive and LM pretraining, data curation has become a first-class component of the recipe and can deliver large gains at fixed compute. Classical coreset and subset-selection methods select training examples to preserve learning dynamics or gradients (Killamsetty et al., 2021a,b; Adiga et al., 2024; Joshi and Mirzasoleiman, 2023; Joshi et al., 2024, 2025a); deduplication and quality filtering scale these ideas to web data (Abbas et al., 2023; Tirumala et al., 2023; Lee et al., 2022; Merrick et al., 2026). Curated open corpora such as DCLM, FineWeb, RedPajama, Dolma, Nemotron-CC, and ÜberWeb have made curation central to LM pretraining (Li et al., 2024a; Penedo et al., 2024; Weber et al., 2024; Soldaini et al., 2024; Su et al., 2024a; Carranza et al., 2026; DatologyAI, 2024b; Maini and DatologyAI Team, 2025), with parallel progress in contrastive image-text training (Gadre et al., 2023; Xu et al., 2023; Fang et al., 2024; DatologyAI, 2024a) and extensions to objectives beyond raw quality such as safety pretraining (Maini et al., 2025). Complementary evidence suggests that much of the leverage for downstream quality lies in the pretraining mixture itself (Baek et al., 2026).

In VLMs, by contrast, prior curation work has mostly studied one axis at a time: multimodal deduplication (Slyman et al., 2024; Liu et al., 2025a), image-text quality filtering (Wang et al., 2024b, 2025b; Chen et al., 2024b), recaptioning of paired data (Chen et al., 2023; Li et al., 2024b; Lai et al., 2023), task-agnostic and task-specific synthetic data (Liu et al., 2024c; Su et al., 2024b; Liao et al., 2025; Joshi et al., 2025b; Yang et al., 2025), mixture design within a single-model recipe (McKinzie et al., 2024; Tong et al., 2024a; Laurençon et al., 2024b; Deitke et al., 2024; Wang et al., 2025a), or releases of curated corpora without a matched-compute reference (Guo et al., 2024; Wiedmann et al., 2025). Recent work has begun composing two of these axes (Liu et al., 2025a; Wang et al., 2025a), but the compositions remain partial and the dominant evaluation pattern reduces the data question to a single benchmark average. Our contribution is to treat pretraining data as the primary variable of interest under matched architecture, recipe, and compute, and to study how curation affects not just benchmark averages but also robustness, generalization beyond benchmarks, and inference-time efficiency.

3 Background

VLM development is commonly partitioned into four phases (Liu et al., 2023b,a; Bai et al., 2023; Wang et al., 2024a; Qwen Team, 2025b; Chen et al., 2024d; Tong et al., 2024a; McKinzie et al., 2024; Deitke et al., 2024; Shi et al., 2024; Laurençon et al., 2024b,a): *VLM pretraining*, which takes a pretrained language model (LM) and a pretrained vision encoder and adapts them into a single joint vision-language model via large-scale image-text training; *supervised / visual instruction fine-tuning (SFT)*, which adapts the pretrained VLM to specific task formats; *preference optimization*, including reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO), which aligns model outputs to human preference data; and *reinforcement learning with verifiable rewards (RLVR)*, which optimizes against verifiable reward signals such as math correctness or code execution. VLM pretraining is the focal phase for this paper: it establishes cross-modal alignment and forms the base on which all downstream stages build, and the pretraining mixture remains where much of the downstream leverage lies (Baek et al., 2026). Every intervention in this paper is confined to pretraining data curation: we run no SFT, no preference optimization, and no RLVR. This makes the comparison against Qwen3-VL and InternVL3.5 a strong one: those public models include all three post-training stages on top of their pretraining, while ours does not.

3.1 Setup

Model. The vision encoder is SigLIP2-SO400M at 384×384 patch-14 resolution (Tschannen et al., 2025), and the language backbone is drawn from the Qwen3 family (Qwen Team, 2025a). We train models at three scales, *1B*, *2B*, and *4B*, each pairing a Qwen3 language-model backbone with SigLIP2-SO400M. High-resolution images are consumed at native resolution via dynamic tiling in the style of InternVL (Chen et al., 2024d): each image is split into an aspect-ratio-aware grid of 384×384 tiles (up to 12 per image) plus a thumbnail, so the total visual-token budget scales with resolution. SigLIP2 encodes each tile into 729 tokens; a 1×3 pixel shuffle then compresses each tile to 243 tokens while expanding feature width from 1,152 to 3,456 channels, and a two-layer MLP projector maps visual features into the language backbone.

Training recipe. All models are trained for 25B tokens in a single stage (projector and LM are trained jointly) at a default context length of 4,096 tokens, with global batch size 512 in 12k steps; sequences are packed to minimize padding. The robustness experiments in §5 sweep the context length to 8k and 16k while keeping the token budget constant. Optimizer, schedule, warmup, weight decay, precision, batch size, and hardware are standard and reported in full in Appendix B.

3.2 Data

***Baseline*: MAmmoTH-VL (single-image subset).** Our *baseline* corpus is the single-image subset of MAmmoTH-VL-12M (Guo et al., 2024), roughly 10M samples after filtering out multi-image instances. MAmmoTH-VL itself aggregates open-source instruction data across natural-image VQA, OCR and document understanding, charts and tables, math and science, captioning, and referring expressions, and supplements them with a re-generation pass in which open VLMs and LMs rewrite annotations into longer, chain-of-thought-style responses; it is broad in coverage and representative of current open VLM training mixtures. We restrict the scope of this study to single-image samples for training, while showing that the resulting curation improvements generalize to multi-image evaluations as well.

DatologyAI-curated mixture. The *curated* mixture is produced by applying the DatologyAI curation pipeline on top of the single-image MAmmoTH-VL subset, matched to the *baseline* in *total training tokens*; we summarize the VLM-specific components below.

Deduplication. MAmmoTH-VL aggregates many open-source datasets, so both images and text are heavily duplicated across sources. We apply multimodal deduplication to reduce redundant examples and improve control over the resulting training mixture.

Mixture design and distribution matching. We analyze the source corpus along several multimodal axes and adjust the resulting mixture to improve coverage, balance, and downstream utility. This includes distribution-level interventions that account for both image and text properties, while preserving broad coverage of the original corpus.

Filtering. We apply multimodal quality filters that operate jointly on image and text signals to remove low-signal, malformed, or off-distribution samples. These filters are tuned to retain diverse and rare-but-useful examples rather than simply maximizing average per-sample quality in isolation.

Synthetic data. Two families of synthetic data complement the filtered, remixed corpus: a *task-agnostic* pipeline that broadens coverage of the source corpus, and *task-specific* generation pipelines for selected capability families.

3.3 Evaluation

We report on two complementary evaluations: **DatBench** (Joshi et al., 2026), the high-fidelity VLM eval suite presented in prior work, and 20 public VLM benchmarks.

DatBench. DatBench draws from over 30 source benchmarks, transforming questions into generative formats where possible, removing samples solvable without the image, and filtering mislabeled or ambiguous examples, to yield roughly 5,000 samples per capability. Coverage spans nine capability axes (grounding, chart, scene, table, spatial, math, counting, document, and general); per-capability scores are averaged into a single summary metric.

IID vs. OOD. IID and OOD are defined with respect to the curation pipeline, not with respect to whether a benchmark is public or to its high-level domain. IID benchmarks correspond to task formats that explicitly informed task-specific synthetic generation or mixture design; OOD benchmarks were not used to shape those curation decisions, even when they probe similar capabilities. BLINK is OOD on a stronger axis: it requires reasoning over multiple images, and our curation is single-image only.

Public benchmark suite. We report on 20 public VLM benchmarks, split into 11 IID and 9 OOD evaluations.

IID public benchmarks cover referring and grounding, general VQA, OCR and document understanding, counting, chart and diagram reasoning, and math: RefCOCO, RefCOCOG, RefCOCO+, PixMo Points, RealWorldQA, TextVQA, DocVQA, CountBench, AI2D, ChartQA, and MathVista.

OOD public benchmarks cover general VQA, OCR and document understanding, captioning, spatial and 3D reasoning, domain-specific recognition, and multi-image reasoning: MMBench, OCRBench, DetailCaps, CAPability, CVBench-2D, CVBench-3D, 3DSRBench, ecommerce brand-ID, and BLINK.

All training data is decontaminated against the full evaluation suite using the procedure described in Appendix A.

4 Data Curation Improves VLMs Across Benchmarks, Domains, and Scales

We report the main result through five complementary views. First, at 2B, curation delivers a large average lift on the 11 IID benchmarks of the public suite. Second, the lift transfers out-of-distribution across nine OOD benchmarks, including domain shift to ecommerce brand-ID and structural shift in BLINK multi-image reasoning despite single-image training. Third, **DatBench** shows the gains span all nine capability axes rather than a narrow subset. Fourth, a grounding deep-dive checks that the headline +57.1pp gain is a bona fide capability shift rather than a scoring artifact. Finally, scaling runs at 1B, 2B, and 4B show the curation effect persists across model size. Full per-benchmark, per-seed numbers are in Appendix C; shared inference settings are in Appendix C.1.

Lift across the 11 IID public-suite benchmarks. Figure 2 reports the curated-vs-baseline comparison on the 11 IID benchmarks in the public suite, spanning referring & grounding, general VQA, OCR & document understanding, counting, chart & diagram reasoning, and math. The 11-IID average rises from 55.1 to 70.5 (+**15.4**pp). Excluding the four referring & grounding evals, the 7-eval IID average still lifts from 68.8 to 73.6 (+**4.8**pp): the headline does not rest on grounding alone, which we examine separately in §4.

Out-of-distribution generalization. The lift transfers beyond the benchmarks the curation pipeline was tuned on. Across 9 OOD benchmarks (general VQA and OCR/document understanding via MMBench and OCRBench, captioning via DetailCaps and CAPability, spatial & 3D reasoning via CVBench-2D, CVBench-3D, and 3DSRBench, domain-specific ecommerce brand-ID, and multi-image reasoning via BLINK), the *curated* model lifts the average by +7.2pp (52.2 → 59.4), and *every* OOD benchmark improves (Figure 3). We single out two cases that sharpen the OOD claim: ecommerce brand-ID, a domain shift, and BLINK, a structural shift from single-image training to multi-image reasoning.

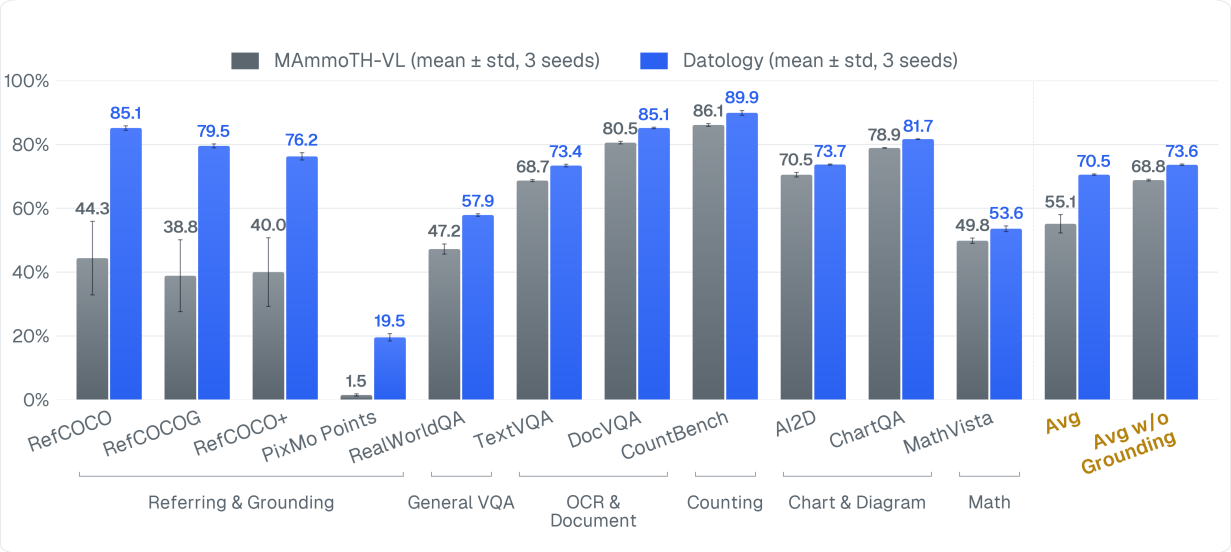


Figure 2 Curation lifts the 11 IID public-suite benchmarks by +15.4pp on average. *Curated* vs. *baseline* on the 11 IID public benchmarks (the 20-eval public suite minus the 9 OOD benchmarks shown in Figure 3) spanning referring & grounding, general VQA, OCR & document understanding, counting, chart & diagram reasoning, and math. Deltas annotated above each pair; the rightmost columns are the 11-IID average (55.1 → 70.5, +15.4pp) and the same average excluding the four referring & grounding evals (68.8 → 73.6, +4.8pp).

Ecommerce brand-ID. The ecommerce brand-ID benchmark, derived from the Shopify product-catalogue dataset (Shopify, 2024), is the cleanest single-domain OOD test in our setup: it targets hyper-specific brand identification from product images, a domain the general-purpose curation pipeline was not designed to address. The *curated* model improves from 31.5 to 36.7 (+5.2pp), evidence that the curation effect transfers well beyond the benchmarks the pipeline was tuned on.

Single-image to multi-image transfer (BLINK). BLINK requires evidence from more than one image, spanning categories such as relative depth and visual correspondence as well as counting and object localization; none of these multi-image comparisons appear in our single-image curation. Aggregate accuracy rises from $41.66 \pm 0.59\%$ (*Baseline*) to $44.75 \pm 0.67\%$ (*Curated*), a gain of +3.09pp, and every *Curated* seed exceeds every *Baseline* seed. At the category level, the largest gains are in `Relative_Depth` (59.1% → 77.4%, +18.3pp) and `Visual_Correspondence` (27.1% → 39.0%, +11.8pp); the latter directly demands cross-image reasoning, which the single-image pipeline was never trained for.

> **Takeaway** | **Data curation generalizes out-of-distribution.** Single-image curation lifts the 9-eval OOD average by **+7.2pp**, improves every OOD benchmark, and transfers to BLINK multi-image reasoning despite no multi-image data appearing in training.

Improvements across all nine DatBench capability axes. Figure 4 reports per-capability DatBench scores at 2B for the MAmmoTH-VL-12M single-image *baseline* and the DatologyAI-curated mixture, averaged over three seeds. Curation raises the 9-capability average from 43.2 to 54.5 (+**11.3**pp), with improvements on every capability axis. The largest gain is on `Grounding` (+57.1pp), followed by `Chart Understanding` (+11.9pp), `Document Understanding` (+9.6pp), `Scene OCR` (+6.7pp), `Diagrams & Tables` (+6.2pp), `Spatial Reasoning` (+4.3pp), `Counting` (+2.5pp), `Math & Logic` (+2.3pp), and `General` (+1.6pp). Excluding Grounding, the remaining 8-capability DatBench average still lifts from 46.3 to 51.9 (+**5.6**pp); the lift is broad rather than grounding-only.

Grounding deep-dive: the gain is bona fide, not a scoring artifact. A +57.1pp gain on a single capability is large enough that the natural question is whether it is real or whether the *curated* model has

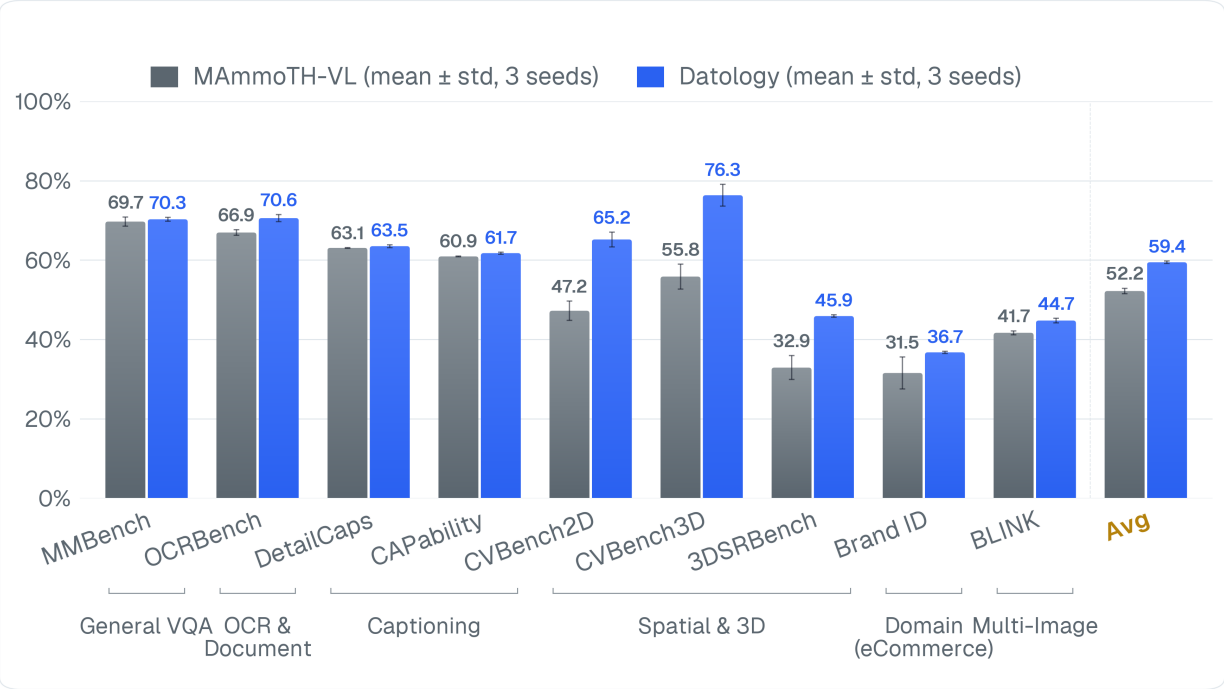


Figure 3 Curation gains transfer out-of-distribution. *Curated* vs. MAmmoTH-VL *baseline* at 2B on 9 out-of-distribution (OOD) public benchmarks spanning general VQA, OCR & document, captioning, spatial & 3D reasoning, domain-specific ecommerce brand-ID, and multi-image reasoning (BLINK). Each bar is averaged over three seeds; error bars show ±1 standard deviation. The rightmost column is the OOD 9-eval average (52.2 → 59.4, +7.2pp). Every OOD benchmark improves under curation, despite none of these benchmarks being used to inform task-specific synthesis or mixture decisions.

found a way to exploit how the metric is computed. RefCOCO-style grounding is convenient here because it is scored against several metrics that probe different facets of the underlying skill: `center_acc` (predicted box centroid lands inside the ground-truth region, probing semantic localization independent of box geometry) and `recall@`k at IoU thresholds 0.3, 0.5, and 0.7 (probing geometric precision, with 0.7 requiring tight box agreement). A scoring-trick gain would show up on one of these metrics and not the others; a bona fide capability gain should move all of them. On the RefCOCO subsets `refcoco_testA`, `refcoco_plus_testA`, and `refcocog_test`, curation wins on every split (Figure 5). Averaged across splits, `center_acc` rises from $69.72 \pm 12.25\%$ to $91.03 \pm 0.40\%$ (+21.3pp), `recall@0.5` from $41.04 \pm 13.69\%$ to $80.29 \pm 1.01\%$ (+39.3pp), and `recall@0.7` from $14.55 \pm 4.85\%$ to $69.96 \pm 0.88\%$ (+55.4pp); `recall@0.3` also moves (full table in Appendix C.6.1). The model identifies the correct referent and localizes it precisely; gains grow at stricter IoU thresholds, ruling out the reading that curation only sharpens coarse localization while leaving fine-grained box regression unchanged. CountBench shows the same pattern across exact, within-1, and within-3 tolerances (Appendix C.6.2).

Gains are strong across model scales. We repeat the matched baseline-vs-curated comparison at 1B, 2B, and 4B parameters, holding training recipe, token budget, and evaluation protocol fixed within each scale. Figure 6 reports average accuracy at each scale on (a) the 20-eval public VLM benchmark suite (IID + OOD union) and (b) **DatBench**. Curation improves accuracy at every scale on both evaluations. On the public suite, the lift is **+13.0pp** at 1B, **+11.7pp** at 2B, and **+14.0pp** at 4B; on **DatBench**, the lift increases from **+8.7pp** at 1B to **+11.3pp** at 2B and **+11.7pp** at 4B. The effect is robust across model size: public-suite gains remain consistently large, while **DatBench** gains increase monotonically with scale.

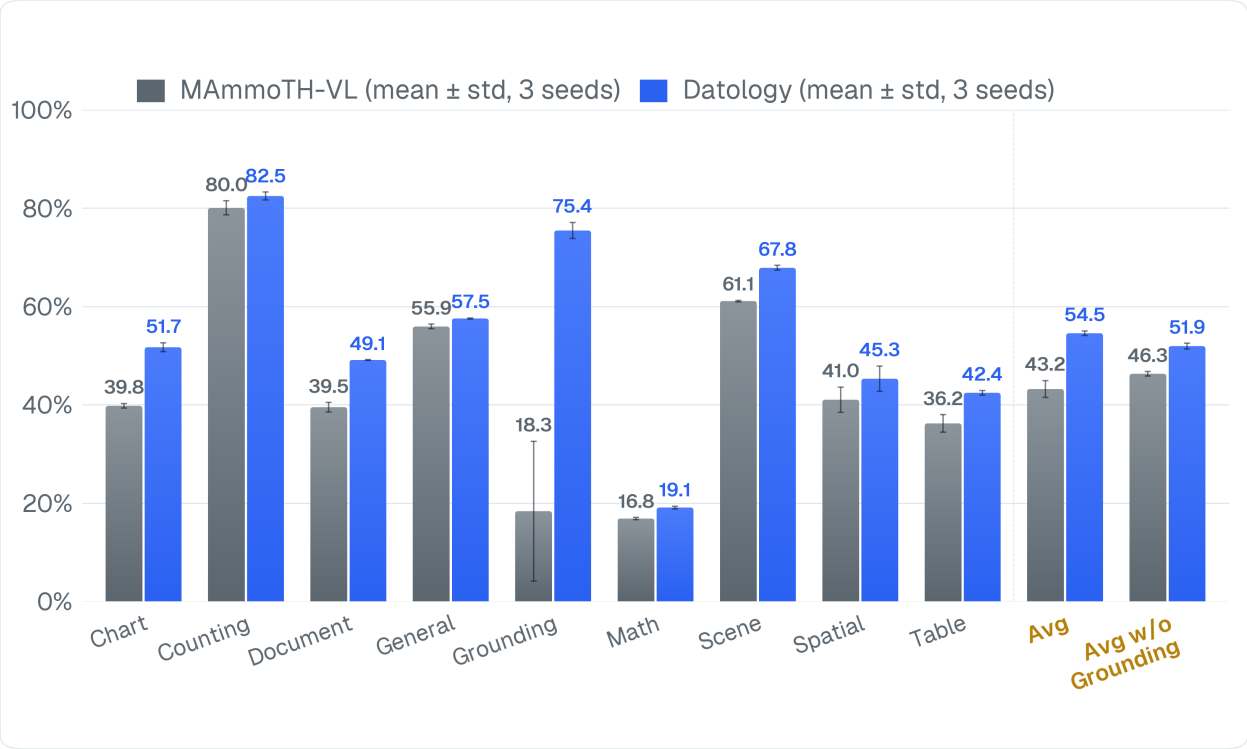


Figure 4 Curation lifts every DatBench capability, not a favored few. Per-capability DatBench scores at 2B, MAmmoTH-VL *baseline* (gray) vs. DatologyAI-curated (blue), each bar averaged over three seeds; error bars show ±1 standard deviation. The rightmost columns are the 9-capability average (43.2 → 54.5, +11.3pp) and the same average excluding Grounding (46.3 → 51.9, +5.6pp). Grounding carries the largest single gain (+57.1pp), and all other capability axes also improve: Chart Understanding (+11.9pp), Document Understanding (+9.6pp), Scene OCR (+6.7pp), Diagrams & Tables (+6.2pp), Spatial Reasoning (+4.3pp), Counting (+2.5pp), Math & Logic (+2.3pp), and General (+1.6pp).

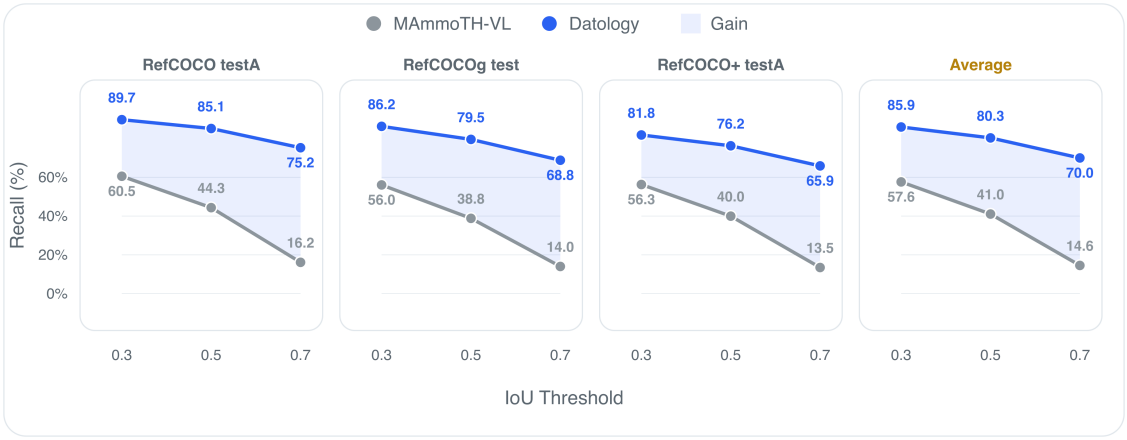
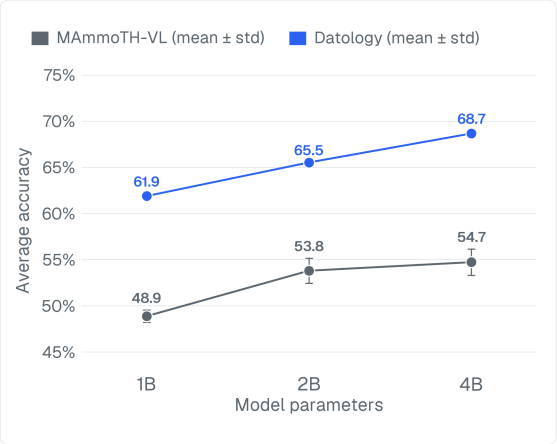
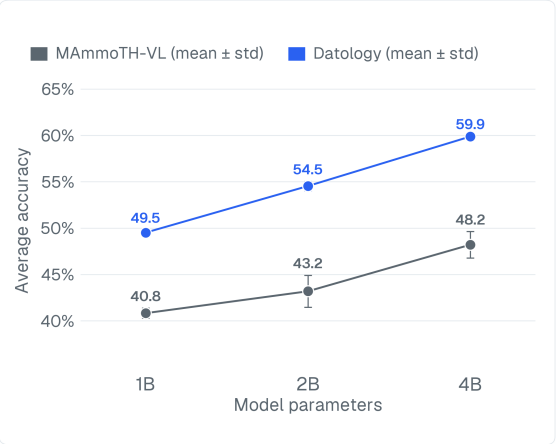


Figure 5 Grounding gains are bona fide: every RefCOCO metric improves under curation, ruling out a scoring artifact. `recall@`k across IoU thresholds and RefCOCO splits, with per-dataset panels and an averaged summary for *Baseline* versus *Curated*.

The cross-scale gain is large enough that the *curated* 1B surpasses the 4× larger *baseline* on both suites: 61.9 **vs** 54.7 on the public suite (+7.2pp) and 49.5 **vs** 48.2 on DatBench (+1.3pp). At a fixed token budget, the 1B uses ∼ 4× less training compute than the 4B *baseline*, so curation closes a 4× scale gap at ∼ 4× less



(a) Average accuracy across the 20-eval public VLM benchmark suite.



(b) Average accuracy across DatBench.

Figure 6 Curation gains are strong across model scales. Average accuracy of the MAmmoTH-VL *baseline* (gray) and DatologyAI-curated model (blue) at 1B, 2B, and 4B parameters, averaged over three seeds; error bars show ±1 standard deviation. Curation improves accuracy at every scale on both (a) the 20-eval public VLM benchmark suite (+13.0, +11.7, and +14.0pp at 1B/2B/4B) and (b) DatBench (+8.7, +11.3, and +11.7pp at 1B/2B/4B).

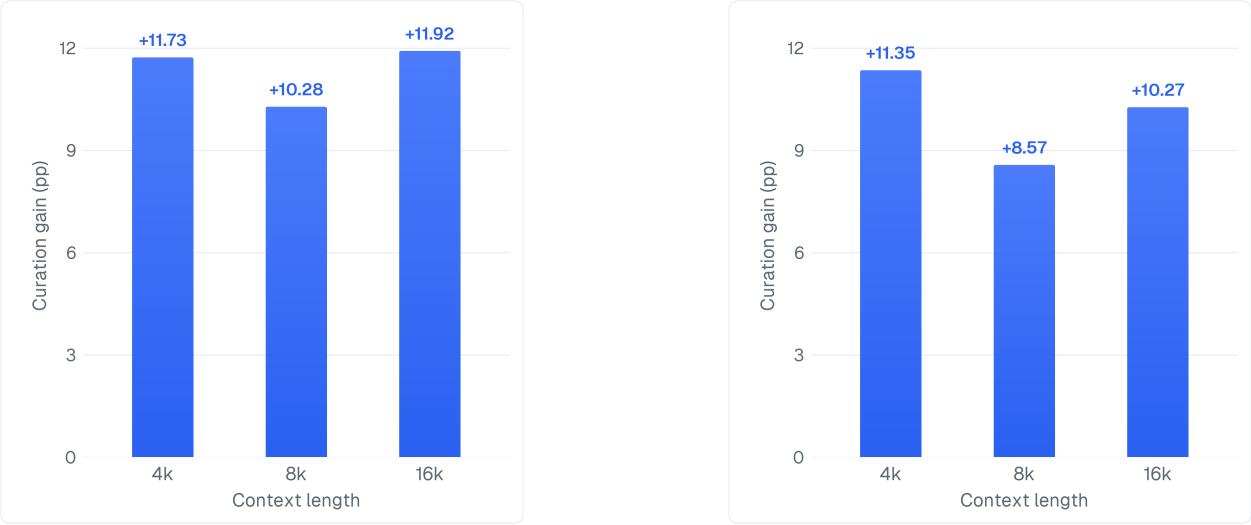
training compute.

> **Takeaway** | **Data curation shifts the cost-quality Pareto frontier across scales.** At fixed architecture, recipe, and compute, it improves every tested scale on both the 20-eval public suite and DatBench, with public-suite gains of +13.0pp, +11.7pp, and +14.0pp at 1B, 2B, and 4B.

5 Data Curation as a Lever for Reliability and Robustness

VLM training exhibits substantial variance across random seeds and sensitivity to hyperparameter choices (Dodge et al., 2020; Bouthillier et al., 2021; Narang et al., 2021), and at the scale of modern pretraining runs, replicating a configuration across multiple seeds or re-running under varied hyperparameters is often infeasible on cost grounds alone. Hyperparameter tuning at scale is itself a major cost driver, so any intervention that reduces sensitivity to seeds and hyperparameters is a candidate lever for lowering the tuning budget needed to land a real gain. We show that data curation moves on both axes: it cuts per-capability variance across training seeds on both evaluation suites, and the *curated* gain is preserved across a context-length sweep from 4k to 16k tokens. Curating the data is itself a tuning-cost intervention.

Reliability: variance reduction across training seeds. We train three seeds per configuration and report per-capability mean ± std on both suites (Figure 8, 4k panel). Averaged across the nine DatBench capabilities at 2B/4k, the cross-seed standard deviation drops from 2.47 to 0.82 pp (−67%) under curation, with most capabilities tightening (Grounding 14.2 → 1.7, Table 1.8 → 0.5, Counting 1.4 → 0.8). Spatial Reasoning is the exception: seed variance stays high under both mixes (2.6 → 2.6), so curation moves the mean on that axis without tightening the noise. A plausible mechanism for the variance reduction is cleaner training signal: filtering, deduplication, and mixture rebalancing remove noisy and ambiguous examples whose gradient contribution shifts unpredictably with seed, leaving a more reproducible learning trajectory. The Spatial Reasoning exception suggests this mechanism is not the whole story, but the broader pattern across capabilities is consistent.

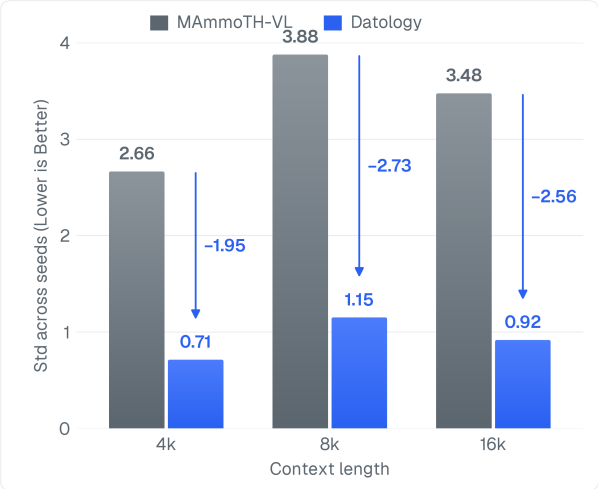


(a) 20-eval public VLM benchmark suite. **(b)** DatBench.

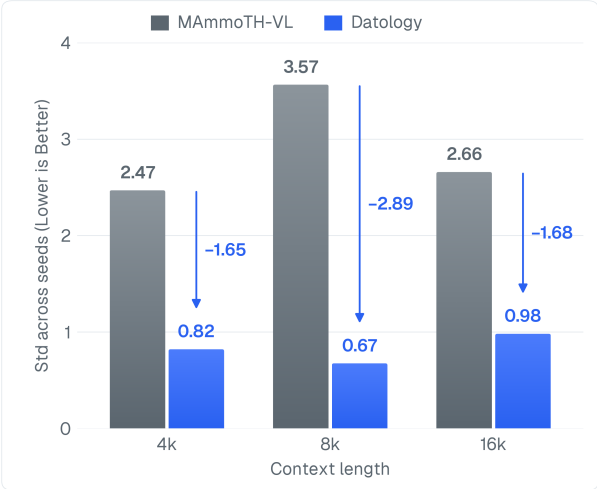
Figure 7 Curation gain persists across context length. Mean curated-vs-baseline accuracy delta (percentage points) at 4k, 8k, and 16k context lengths on (a) the 20-eval public VLM benchmark suite (+11.7, +10.3, +11.9pp) and (b) DatBench (+11.3, +8.6, +10.3pp). The gap holds at every context length we probed; no setting closes it.

Robustness: consistency across context-length variation. We sweep context length over 4k, 8k, and 16k tokens, a hyperparameter that varies widely across model releases and training setups, and ask whether both the mean lift and the variance reduction survive at each setting. They do, on both axes. Figure 7 reports the curated-vs-baseline mean delta at each context length: on the 20-eval public suite the gain is +11.7pp at 4k, +10.3pp at 8k, and +11.9pp at 16k; on DatBench it is +11.3pp at 4k, +8.6pp at 8k, and +10.3pp at 16k. No context length closes the gap. Figure 8 reports the corresponding mean per-capability cross-seed standard deviation: on the public suite curation cuts std from 2.66 → 0.71 at 4k, 3.88 → 1.15 at 8k, and 3.48 → 0.92 at 16k; on DatBench from 2.47 → 0.82 at 4k, 3.57 → 0.67 at 8k, and 2.66 → 0.98 at 16k. The reliability gain at the default 4k setting is not co-adapted to that setting; curation tightens cross-seed std at every context length we probed. Full per-seed numbers and the context-length sweep table are in Appendix C; broader hyperparameter sweeps (optimizer, schedule, batch size) remain future work.

> **Takeaway** | **Data curation is reliable across training seeds and robust across context lengths.** Curation reduces average per-capability cross-seed standard deviation from 2.47 to 0.82pp on DatBench and preserves its accuracy lift across 4k, 8k, and 16k context lengths.



(a) 20-eval public VLM benchmark suite.



(b) DatBench.

Figure 8 Variance reduction persists across context length. Mean per-capability cross-seed standard deviation (percentage points, lower is better) for MAmmoTH-VL *baseline* (gray) vs. DatologyAI-curated (blue) at 4k, 8k, and 16k context lengths on (a) the 20-eval public VLM benchmark suite (2.66 → 0.71, 3.88 → 1.15, 3.48 → 0.92) and (b) DatBench (2.47 → 0.82, 3.57 → 0.67, 2.66 → 0.98). The reliability gain at the default 4k context is not co-adapted to that setting; curation tightens cross-seed std at every context length we probed.

6 Beyond Benchmarks: Data Curation Generalizes to Real-World Use

Benchmark scores capture a fraction of what users notice. The capability surface of these models is jagged (Dell'Acqua et al., 2023; Tong et al., 2024b): a model can score well in aggregate while failing on the queries a user actually issues. Behaviors that matter in practice, such as reading a sign, asking about an unfamiliar object, or expecting a concise answer, are unevenly captured by standard benchmarks. The question for pretraining-data curation is whether gains on scored benchmarks transfer to these user-facing behaviors when the model is queried freely, on images and prompts it has not been optimized against.

We compare the *curated* 2B model against the compute-matched MAmmoTH-VL *baseline* and Qwen3-VL-2B (Qwen Team, 2025b) as a frontier 2B reference on four target behaviors: *honesty* (refusing to confabulate when an attribute is absent), *specificity* (giving a specific identification rather than a categorical description), *conciseness* (answering with the requested information without unrequested elaboration), and *non-refusal* (engaging with benign prompts rather than refusing them). Figure 9 depicts a representative example for each behavior.

Methodology. We had frontier multimodal VLM agents query the three models across roughly 1,100 single-image prompts spanning OCR, brand and franchise recognition, multi-element scene description, attribute queries on present and absent properties, and refusal calibration; the agents proposed images and prompts, collected outputs, and triaged them. We adopt this approach because the coverage it enables across image and prompt types is unreachable by manual authoring, and we mitigate its known noise (agent-proposed prompts drifting toward agent-legible distributions, classification noise compounding at scale) by manually verifying every claim that appears in this section. Inference settings match §4.

Selected wins versus the matched-compute *baseline*: specificity and honesty. Specificity is the cleanest signal. Across n=67 knowledge-grounded recognition queries (cake franchise, vehicle make from emblem, brand identification, bird species, concrete multi-element scene description), the *curated* model strictly outperforms the *baseline* on 44 cases and is strictly outperformed on 0; every disagreement is a *curated*

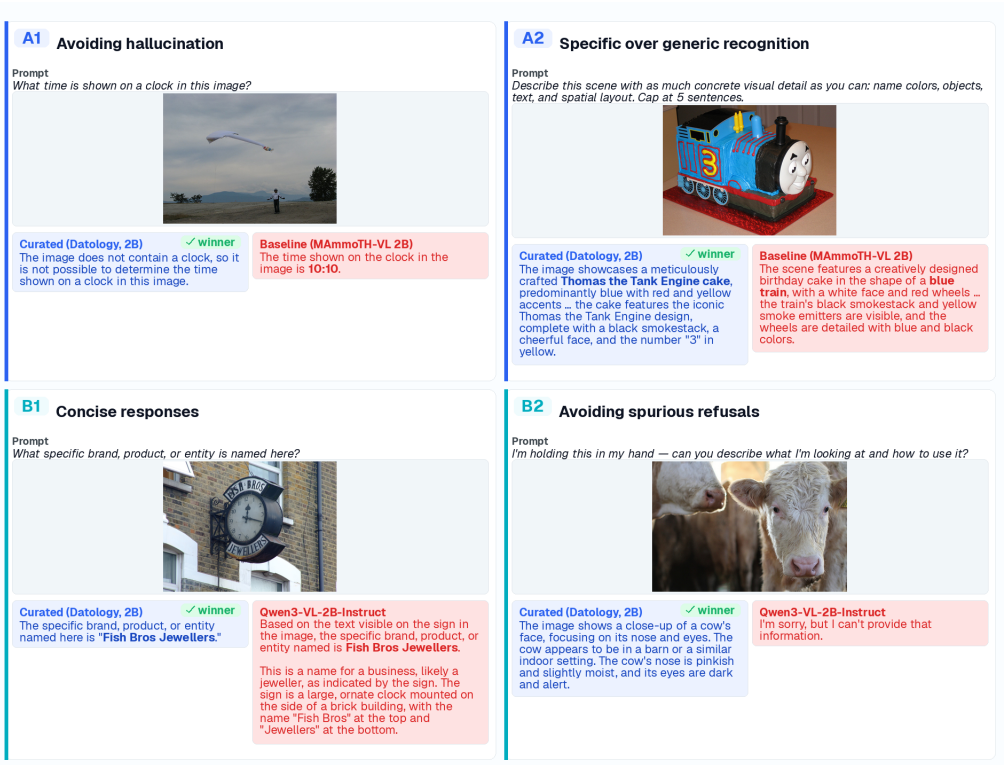


Figure 9 The *curated* model answers more honestly and more specifically than the matched-compute *baseline*, and more concisely and refuses fewer benign prompts than a frontier 2B reference. Four selected cases (*curated* wins each). *Top row, curated > matched-compute baseline:* A1 *honesty* on a no-clock scene; *baseline* invents *10:10*, the canonical watch-ad time, on a kite-flying image that contains no clock. A2 *specificity*: *curated* names the *Thomas the Tank Engine* cake and reads the "3" on the smokestack, while *baseline* returns a generic *blue train cake* and fabricates yellow smoke emitters and blue-and-black wheels on an open "describe this scene" prompt that does not leak the franchise. *Bottom row, curated > Qwen3-VL-2B:* B1 *conciseness*: both models correctly read *Fish Bros Jewellers* from an ornate street clock; *curated* answers in one sentence while Qwen3-VL appends two paragraphs of unrequested description about the sign and building. B2 *non-refusal* on a benign close-up of two cows with a prompt asking what's in the user's hand and how to use it: *curated* narrates the scene; Qwen3-VL-2B returns *I'm sorry, but I can't provide that information.* All outputs verbatim at $T = 0$.

win. Figure 9, panel A2 shows a representative case: on an open "describe this scene" prompt that does not leak the franchise, the *curated* model names a *Thomas the Tank Engine* cake and reads the number "3" off the smokestack, while the *baseline* returns a generic *blue train cake* and fabricates yellow smoke emitters and blue-and-black wheels. On honesty, the qualitative pattern is clearer than the aggregate. Across the n=15 absent-attribute queries the agents surfaced (a thin sample, so individual cases anchor the result), the *curated* model declines or hedges correctly on 7 **wins** versus 4 losses with 4 ties: on a kite-flying scene with no clock present, the *baseline* invents *10:10*, the canonical watch-ad time and a training-data artifact rather than a perceptual error (Figure 9, panel A1); on a separate query asking about a purple hat in a scene that contains none, the *baseline* asserts one regardless.

Selected wins versus a frontier 2B reference: conciseness and non-refusal. The *curated* model answers more concisely and refuses fewer benign queries than Qwen3-VL-2B (Figure 9, bottom row). On the n=289 queries where the *curated* 2B and Qwen3-VL-2B converge on the same answer ($\geq$3 shared content words), the *curated* model's median response length is 75 **characters** versus 364 **characters** for Qwen3-VL-2B (nearly 5$\times$ shorter at equivalent correctness); Qwen3-VL-2B's `max_tokens` cap is hit on 20.1% of queries against 4.9% for the *curated* model. The conciseness gap holds across the entire response-length distribution, not just at the median or the cap. On a brand-recognition prompt over an ornate street clock, both models correctly read *Fish Bros Jewellers*; the *curated* answer is one sentence while Qwen3-VL appends two paragraphs

of unrequested description about the sign and building (panel B1). For non-refusal, across the full n=358 three-way query set Qwen3-VL-2B returned its blanket *"I'm sorry, but I can't provide that information"* on 30 queries (8.4%); on five of these (franchise identification, breed classification, object close-ups), the *curated* 2B answers correctly. On a benign close-up of two cows with the prompt "describe what's in the user's hand and how to use it," the *curated* model narrates the scene while Qwen3-VL-2B returns *I'm sorry, but I can't provide that information* (panel B2).

> **Takeaway** | **Data curation improves reliability across training seeds and robustness across context lengths.** It cuts average per-capability cross-seed standard deviation from 2.47 to 0.82pp on DatBench and preserves its accuracy lift across 4k, 8k, and 16k context lengths.

7 Data Curation Is Pareto-Dominant on Inference Cost at Every Scale

As VLMs move into production, inference cost matters: at scale, even modest increases in generated tokens or active parameters translate into large serving costs. We evaluate whether pretraining data curation improves both model quality and the cost of achieving it at inference time. Across all tested scales, our *curated* models are simultaneously more accurate and cheaper to run than the pretraining-compute-matched baselines, and reach competitive accuracy against heavily post-trained public models at substantially lower response FLOPs.

Setup. We evaluate inference-time efficiency on the full 20-eval public benchmark suite from §4. For each model we report *average accuracy* across the evals against a *response-FLOPs proxy*: 2 × active params × mean generated tokens per response (the standard transformer decode-only forward-pass cost; full derivation and per-model accounting in Appendix C.4, Table 8), computed from raw outputs using each model's tokenizer. Lower response FLOPs at equal accuracy is a more usable model at fixed inference budget. We compare the *curated* 1B, 2B, and 4B against the matched-compute MAmmoTH-VL baselines at the same scales, and against Qwen3-VL-2B/4B (Qwen Team, 2025b), InternVL3-2B, InternVL3.5-2B/4B (Chen et al., 2024d), and Qwen3.5-2B as frontier 2B/4B references.[2] The comparison isolates two effects: at fixed scale, *curated* models reach higher accuracy at lower response FLOPs than their compute-matched baselines, and across scales they match or approach extensively post-trained references at substantially lower inference cost.

Comparison to compute-matched baselines. Against the matched-compute MAmmoTH-VL *baseline*, our curated dataset is Pareto-dominant at every scale we tested, simultaneously raising accuracy and lowering response FLOPs (Figure 10): at 1B, accuracy rises from 48.9% to 61.9% (+13.0pp) while response FLOPs drop from 1.15×10^{11} to 6.48×10^{10} ($\sim$ 44% lower); at 2B, 53.8% → 65.5% (+11.7pp) and 2.34×10^{11} → 1.77×10^{11} ($\sim$ 24% lower); at 4B, 54.7% → 68.7% (+14.0pp) and 5.24×10^{11} → 2.86×10^{11} ($\sim$ 45% lower).

Comparison to public frontier models. Against extensively post-trained public references the *curated* model is Pareto-dominant on response FLOPs at near-frontier accuracy (Figure 10). At 4B, the *curated* model reaches 68.7% average accuracy at 2.86×10^{11} FLOPs, against Qwen3-VL-4B's 71.8% at 9.36×10^{11}: Qwen3-VL gains +3.1pp accuracy at 3.3× the inference cost. At 2B, the *curated* model reaches 65.5% at 1.77×10^{11} against Qwen3-VL-2B's 67.3% at 4.27×10^{11}: a 1.8pp gap at 2.4× lower inference cost. The *curated* 4B is within 0.7pp of Qwen3.5-2B (69.4% at 6.77×10^{11}) at 2.4× lower inference cost, and exceeds Qwen3-VL-2B's accuracy (67.3% at 4.27×10^{11}) by 1.4pp at 1.5× lower inference cost.

[2] Qwen3.5-4B is omitted from the inference-efficiency Pareto: its mean response-token count ($\sim$ 1284, Table 7) is roughly an order of magnitude above all other comparators and would compress the response-FLOPs axis into illegibility. Its training compute is included in the training-FLOPs comparisons elsewhere in the paper.

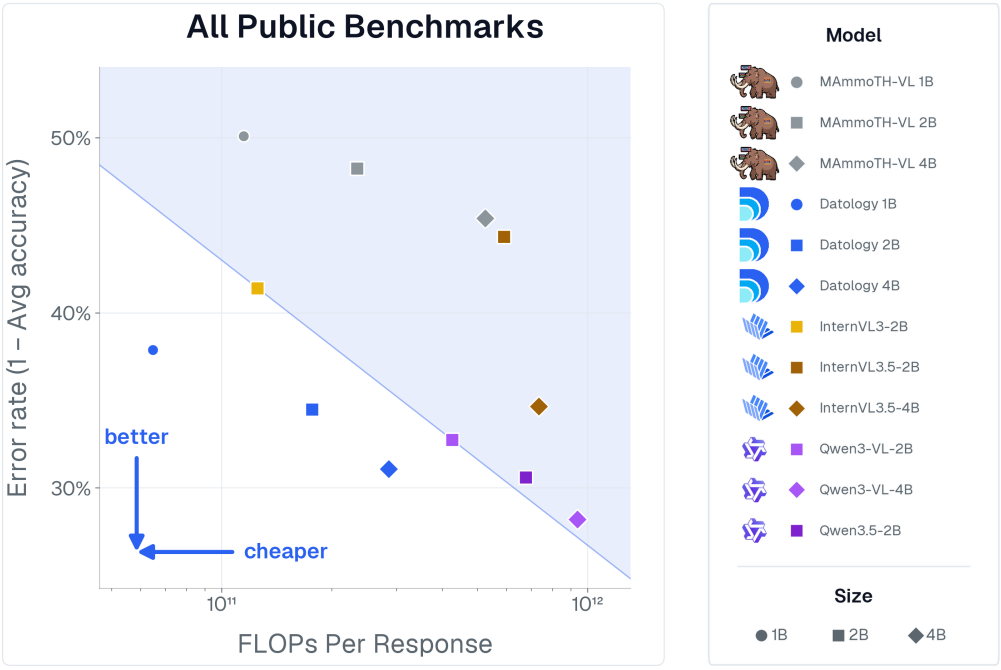


Figure 10 Curation is Pareto-dominant on inference cost at every scale we tested. Average accuracy across the 20-eval public suite (Table 5) against the response-FLOPs proxy ($2 \cdot$ active params $\cdot$ mean generated tokens, log scale, lower is better; lower-left is better overall). Marker shape encodes parameter scale (circle = 1B, square = 2B, diamond = 4B); marker color encodes the model family (blue = *curated*, gray = MAmmoTH-VL *baseline*, other colors = post-trained public references). Same-shape markers compare different curation strategies at fixed scale; same-color markers compare scales within a family. The diagonal traces the Pareto frontier defined by the matched-compute MAmmoTH-VL baselines, and every *curated* point lies below-left of it. DatologyAI-curated 1B/2B/4B (blue) Pareto-dominate the matched-compute MAmmoTH-VL baselines (gray) at every scale, and reach near-frontier accuracy at 1.5–3.3× lower response cost than Qwen3-VL-2B/4B. The *curated* 1B sits furthest into the lower-left because it is both smaller (lower active-params factor) and produces shorter responses than the larger *curated* models (lower mean-tokens factor). Outputs sampled at $T = 0$.

> **Takeaway** | **Data curation is Pareto-dominant on inference cost.** At every tested scale, *curated* models raise accuracy while lowering response FLOPs versus matched-compute baselines; at 4B, the *curated* model reaches near-frontier accuracy at 3.3× lower response FLOPs than Qwen3-VL-4B.

8 Conclusion

All five properties posed in §1 hold under matched compute. The curation lift is *large*: a single curation pass on top of the MAmmoTH-VL single-image subset produces +11.7pp on 20 public VLM benchmarks, +11.3pp across DatBench's nine capability axes, and closes the cost-quality gap to extensively post-trained references at up to $\sim$ 150× less training compute. It is *OOD-generalizing*: single-image curation lifts multi-image reasoning on BLINK by +3.09pp overall, with Visual Correspondence gaining +11.8pp despite demanding cross-image reasoning, and improves every OOD benchmark, including hyper-specific domains such as ecommerce brand identification that the pipeline was not tuned on. It is *reliable*: mean per-capability standard deviation across training seeds drops by $\sim$ 67%, and the curated-vs-baseline delta survives context-length variation from 4k to 16k. It *manifests beyond benchmarks*: on roughly 1,100 open-ended queries the *curated* 2B is more honest and more specific than the matched-compute *baseline*, and answers more concisely and refuses fewer benign queries than a frontier 2B reference. It is *inference-efficient*: at every

scale tested (1B, 2B, 4B), our curation pipeline yields models that simultaneously raise accuracy and lower response FLOPs against the matched-compute *baseline*, with the *curated* 4B reaching near-frontier accuracy at 3.3× lower response-FLOPs cost than Qwen3-VL-4B. Pretraining data curation alone, holding architecture, recipe, and compute fixed, delivers VLM gains that are large, OOD-generalizing, reliable, evident beyond benchmarks, and inference-efficient, at up to ∼ 150× less training compute than extensively post-trained references. Treating data as a first-class design variable is a high-leverage path to better VLMs.

9 Contributions and Acknowledgements

Core Contributors Siddharth Joshi, Haoli Yin, Rishabh Adiga, Haakon Mongstad, and Alvin Deng.

for tiling the data, aligning the modalities, and grounding every claim in this paper.

Technical Contributors Aldo Carranza, Alex Fang, Amro Abbas, Anshuman Suri, Brett Larsen, Daniel Zayas, Darren Teh, David Schwab, Diego Kiner, Fan Pan, Jack Urbanek, Jason Lee, Jason Telanoff, Josh Wills, Kaleigh Mentzer, Luke Merrick, Maximilian Böther, Parth Doshi, Paul Burstein, Pratyush Maini, Ties Robroek, Tony Jiang, Vidhi Jain, Vineeth Dorna, and Zhengping Wang.

the wide-receptive-field ensemble that filtered the noise, synthesized the signal, and ablated every datum of the pipeline.

Leadership Bogdan Gaza, Ari Morcos, and Matthew Leavitt.

the ground-truth supervision that kept cross-modal alignment on track and prevented collective mode collapse.

Acknowledgements Liz Gatapia (*for incredible logo design*), Dan Darnell, Elise Clark, Jacqueline Liu, Janelle Raymundo, Kylie Clement, Sama Iqbal, Sylvia Hoang, Tiffanie Pham, and Zeek Politzer.

the human-in-the-loop supervision that kept the team well-regularized and the signal flowing off-screen.

References

A. Abbas, K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

M. Acharya, K. Kafle, and C. Kanan. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

R. Adiga, L. Subramanian, and V. Chandrasekaran. Designing informative metrics for few-shot example selection. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10127–10135, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.602. URL https://aclanthology.org/2024.findings-acl.602/.

R. Adiga, B. Nushi, and V. Chandrasekaran. Attention speaks volumes: Localizing and mitigating bias in language models. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26403–26423, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1281. URL https://aclanthology.org/2025.acl-long.1281/.

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

C. Baek, R. P. Monti, D. Schwab, A. Abbas, R. Adiga, C. Blakeney, M. Böther, et al. The finetuner's fallacy: When to pretrain with your finetuning data. *arXiv preprint arXiv:2603.16177*, 2026.

J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Sepah, E. Raff, K. Madan, V. Voleti, S. E. Kahou, V. Michalski, D. Serdyuk, T. Arbel, C. Pal, G. Varoquaux, and P. Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

A. G. Carranza, K. Mentzer, R. P. Monti, A. Fang, A. Deng, A. Abbas, A. Suri, et al. Überweb: Insights from multilingual curation for a 20-trillion-token dataset. *arXiv preprint arXiv:2602.15210*, 2026.

L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. ShareGPT4V: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, and F. Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

R. Chen, Y. Wu, L. Chen, G. Liu, Q. He, T. Xiong, C. Liu, J. Guo, and H. Huang. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection, 2024b. URL https://arxiv.org/abs/2402.12501.

Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024c.

Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024d.

W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

W. Dai, N. Lee, B. Wang, Z. Yang, Z. Liu, J. Barker, T. Rintamaki, M. Shoeybi, B. Catanzaro, and W. Ping. NVLM: Open frontier-class multimodal LLMs. *arXiv preprint arXiv:2409.11402*, 2024.

DatologyAI. CLIP gets a data upgrade: Outperforming SoTA with improved data curation only, 2024a. URL https://www.datologyai.com/blog/clip-gets-a-data-upgrade-outperforming-sota-with-improved-data-curation-only.

DatologyAI. Technical deep-dive: Curating our way to a state-of-the-art text dataset, 2024b. URL https://www.datologyai.com/blog/technical-deep-dive-curating-our-way-to-a-state-of-the-art-text-dataset.

M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

F. Dell'Acqua, E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Working Paper 24-013, 2023.

J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. In *arXiv preprint arXiv:2002.06305*, 2020.

A. Fang, A. M. Jose, A. Jain, L. Schmidt, A. Toshev, and V. Shankar. Data filtering networks. In *International Conference on Learning Representations (ICLR)*, 2024.

C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. BLINK: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

J. Guo, T. Zheng, Y. Bai, B. Li, Y. Wang, K. Zhu, Y. Li, G. Neubig, W. Chen, and X. Yue. MAmmoTH-VL: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024. URL https://arxiv.org/abs/2412.05237.

W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Y. Hu, and S. Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

S. Joshi and B. Mirzasoleiman. Data-efficient contrastive self-supervised learning: Most beneficial examples for supervised learning contribute the least. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 15356–15370. PMLR, 2023. URL https://proceedings.mlr.press/v202/joshi23b.html.

S. Joshi, A. Jain, A. Payani, and B. Mirzasoleiman. Data-efficient contrastive language-image pretraining: Prioritizing data quality over quantity. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v238/.

S. Joshi, J. Ni, and B. Mirzasoleiman. Dataset distillation via knowledge distillation: Towards efficient self-supervised pre-training of deep networks. *International Conference on Learning Representations (ICLR)*, 2025a.

S. Joshi, B. Nushi, V. Balachandran, V. Chandrasekaran, V. Vineet, N. Joshi, and B. Mirzasoleiman. MM-Gen: Enhancing task performance through targeted multimodal data curation. *arXiv preprint arXiv:2501.04155*, 2025b.

S. Joshi, H. Yin, R. Adiga, R. P. Monti, A. G. Carranza, A. Fang, A. Deng, A. Abbas, et al. Datbench: Discriminative, faithful, and efficient vlm evaluations. *arXiv preprint arXiv:2601.02316*, 2026.

S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, and R. Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021a.

K. Killamsetty, X. Zhao, F. Chen, and R. Iyer. RETRIEVE: Coreset selection for efficient and robust semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Z. Lai, H. Zhang, B. Zhang, W. Wu, H. Bai, A. Timofeev, X. Du, Z. Gan, J. Shan, C.-N. Chuah, Y. Yang, and M. Cao. VeCLIP: Improving CLIP training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*, 2023.

H. Laurençon, A. Marafioti, V. Sanh, and L. Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024a.

H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024b.

K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023b.

J. Li, A. Fang, G. Smyrnis, M. Ivgi, M. Jordan, S. Gadre, H. Bansal, E. Guha, S. Keh, K. Arora, et al. DataComp-LM: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024a.

L. Li, Z. Xie, M. Li, S. Chen, P. Wang, L. Chen, Y. Yang, B. Wang, and L. Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023c.

X. Li, H. Tu, M. Hui, Z. Wang, B. Zhao, J. Xiao, S. Ren, J. Mei, Q. Liu, H. Zheng, Y. Zhou, and C. Xie. What if we recaption billions of web images with LLaMA-3? *arXiv preprint arXiv:2406.08478*, 2024b.

Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023d.

X. Liao et al. Unicorn: Text-only data synthesis for vision language model training, 2025. URL https://arxiv.org/abs/2503.22655.

H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

J. Liu et al. Quality over quantity: Boosting data efficiency through ensembled multimodal data curation, 2025a. URL https://arxiv.org/abs/2502.08211.

Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023c.

Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 2024b.

Z. Liu, Z. Sun, Y. Zang, X. Dong, Y. Cao, H. Duan, D. Lin, and J. Wang. Visual-RFT: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025b.

Z. Liu et al. SynthVLM: High-efficiency and high-quality synthetic data for vision-language models, 2024c. URL https://arxiv.org/abs/2407.20756.

P. Maini and DatologyAI Team. BeyondWeb: Lessons from scaling synthetic data for trillion-scale pretraining. *arXiv preprint arXiv:2508.10975*, 2025. URL https://www.datologyai.com/blog/beyondweb.

P. Maini, S. Goyal, D. Sam, A. Robey, Y. Savani, Y. Jiang, A. Zou, M. Fredrikson, Z. C. Lipton, and J. Z. Kolter. Safety pretraining: Toward the next generation of safe ai, 2025. URL https://arxiv.org/abs/2504.16980.

M. Mathew, D. Karatzas, and C. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers, et al. MM1: Methods, analysis & insights from multimodal LLM pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

F. Meng, L. Du, Z. Liu, Z. Zhou, Q. Lu, D. Fu, B. Shi, W. Wang, J. He, K. Zhang, et al. MM-Eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

L. Merrick, A. Fang, A. G. Carranza, A. Deng, A. Abbas, B. Larsen, C. Blakeney, et al. Luxical: High-speed lexical-dense text embeddings. DatologyAI technical report, 2026. Lexical-dense text embeddings used for large-scale data filtering.

S. Narang, H. W. Chung, Y. Tay, W. Fedus, T. Fevry, M. Matena, K. Malkan, N. Fedus, D. Bahri, T. Schuster, H. S. Zheng, N. Houlsby, and D. Metzler. Do transformer modifications transfer across implementations and applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, and T. Dekel. Teaching CLIP to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

G. Penedo, H. Kydlíček, L. B. allal, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, and T. Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Qwen Team. Qwen3-VL, 2025b. URL https://qwenlm.github.io/blog/qwen3-vl/.

Qwen Team. Qwen3.5, 2026. URL https://qwen.ai/blog?id=qwen3.5.

H. Shen, P. Liu, J. Li, C. Fang, Y. Ma, J. Liao, Q. Shen, Z. Zhang, K. Zhao, Q. Zhang, et al. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

M. Shi, F. Liu, S. Wang, S. Liao, S. Radhakrishnan, D.-A. Huang, H. Yin, K. Sapra, Y. Yacoob, H. Shi, et al. Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024.

Shopify. Product catalogue. https://huggingface.co/datasets/Shopify/product-catalogue, 2024. Hugging Face dataset.

E. Slyman, S. Lee, M. Kahng, and A. T. Kim. FairDeDup: Detecting and mitigating vision-language fairness disparities in semantic dataset deduplication, 2024. URL https://arxiv.org/abs/2404.16123.

L. Soldaini, R. Kinney, A. Bhagia, D. Schwenk, D. Atkinson, R. Authur, B. Bogin, K. Chandu, J. Dumas, Y. Elazar, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

D. Su, K. Kong, Y. Lin, J. Jennings, B. Norick, M. Kliegl, M. Patwary, M. Shoeybi, and B. Catanzaro. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*, 2024a.

X. Su et al. SK-VQA: Synthetic knowledge generation at scale for training context-augmented multimodal llms, 2024b. URL https://arxiv.org/abs/2406.19593.

Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, et al. Aligning large multimodal models with factually augmented RLHF. *arXiv preprint arXiv:2309.14525*, 2023.

K. Tirumala, D. Simig, A. Aghajanyan, and A. S. Morcos. D4: Improving LLM pretraining via document de-duplication and diversification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

M. Tschannen et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

W. Wang, K. Mrini, L. Yang, S. Kumar, Y. Tian, X. Yan, and H. Wang. Finetuned multimodal language models are high-quality image-text data filters, 2024b. URL https://arxiv.org/abs/2403.02677.

W. Wang et al. Open-Qwen2VL: Compute-efficient pre-training of fully-open multimodal llms on academic resources, 2025a. URL https://arxiv.org/abs/2504.00595.

W. Wang et al. UniFilter: A unified multimodal quality classifier for filtering vision-language pretraining data, 2025b. URL https://arxiv.org/abs/2510.15162.

M. Weber, D. Y. Fu, Q. Anthony, Y. Oren, S. Adams, A. Alexandrov, X. Lyu, H. Nguyen, X. Yao, V. Adams, et al. RedPajama: an open dataset for training large language models. *arXiv preprint arXiv:2411.12372*, 2024.

L. Wiedmann, O. Zohar, A. Mahla, X. Wang, R. Li, T. Frere, L. von Werra, A. Roy Gosthipaty, and A. Marafioti. FineVision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025. URL https://huggingface.co/datasets/HuggingFaceM4/FineVision.

xAI. Grok-1.5 vision preview (RealWorldQA), 2024. URL https://x.ai/news/grok-1.5v.

H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer. Demystifying CLIP data. *arXiv preprint arXiv:2309.16671*, 2023.

Y. Yang et al. Scaling text-rich image understanding via code-guided synthetic multimodal data generation, 2025. URL https://arxiv.org/abs/2502.14846.

L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, 2016.

T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun, and T.-S. Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Y. Zhou, C. Cui, R. Rafailov, C. Finn, and H. Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

A Multimodal Decontamination Pipeline

The headline gains reported in §4 are a real capability lift only if *curated* and *baseline* models alike saw no eval samples during training. We decontaminate every training corpus against the full evaluation suite before any model is trained, so the curated-vs-baseline deltas reported throughout the paper are deltas between equally decontaminated runs. This appendix formalizes what multimodal decontamination should mean and describes the pipeline that implements it.

Multimodal decontamination is subtle because image and text signals are each *individually* misleading indicators of leakage. The same image is frequently re-annotated across datasets with different questions and answers (legitimate training signal that image-only deduplication would wrongly discard); conversely, template questions such as *"What is the title of this chart?"* appear against thousands of unrelated images (text-only matching flags generic templates as contamination). We therefore build the pipeline around the principle that a training sample constitutes leakage only when *both* its image and its text substantially match an eval sample. The pipeline filters training multimodal data against eval multimodal data; it cannot filter what the LM backbone may have memorized during its own pretraining, a symmetric risk we acknowledge but do not address.

Formally, a training document x_t is contaminated with respect to an eval document x_e iff

$$\mathrm{sim}_{\mathrm{img}}(x_t, x_e) \geq \tau_I \quad \text{and} \quad C_{\mathrm{text}}(x_e \to x_t) \geq \tau_T, \tag{A.1}$$

where $\mathrm{sim}_{\mathrm{img}}$ is cosine similarity between DINOv2 ViT-B/14 image embeddings (Oquab et al., 2023) and C_{text} is the directional n-gram containment defined below. A training document is flagged for removal if Eq. A.1 holds for at least one x_e across the union of all evaluation samples. Table 1 enumerates the three image/text regimes and the rationale for the joint criterion; Figure 11 shows representative flagged and non-flagged pairs from each.

Image match	Text match	Decision	Rationale
✓	✓	**Remove**	model can memorize the eval answer
✓	×	Keep	same image, different Q/A, thus valid training signal
×	✓	Keep	template Q ("what is the title?") appears on every chart

Table 1 Only the joint image-and-text criterion removes genuine leakage without discarding legitimate training data. For partial text matches, the threshold τ_T on C_{text} resolves the boundary between the first and second rows.

Algorithm 1 Two-stage multimodal decontamination cascade.

```
for each training document x_t do
    sim_img ← max DINOv2 cosine similarity between x_t and any eval image
    if sim_img ≥ τ_I then                                  ▷ Stage 1: image recall
        for each eval document x_e do
            C_text ← one-way n-gram containment of x_e's Q+A in x_t's Q+A
            if C_text ≥ τ_T then                           ▷ Stage 2: text precision
                remove x_t and break
            end if
        end for
    end if
end for
```

We implement Eq. A.1 as an image-recall + text-precision cascade (Algorithm 1). Stage 1 embeds every image with DINOv2 and aggregates similarity to parent documents via MAX, so a document inherits the similarity of its most similar constituent image. Stage 2, for every document above τ_I, computes text containment

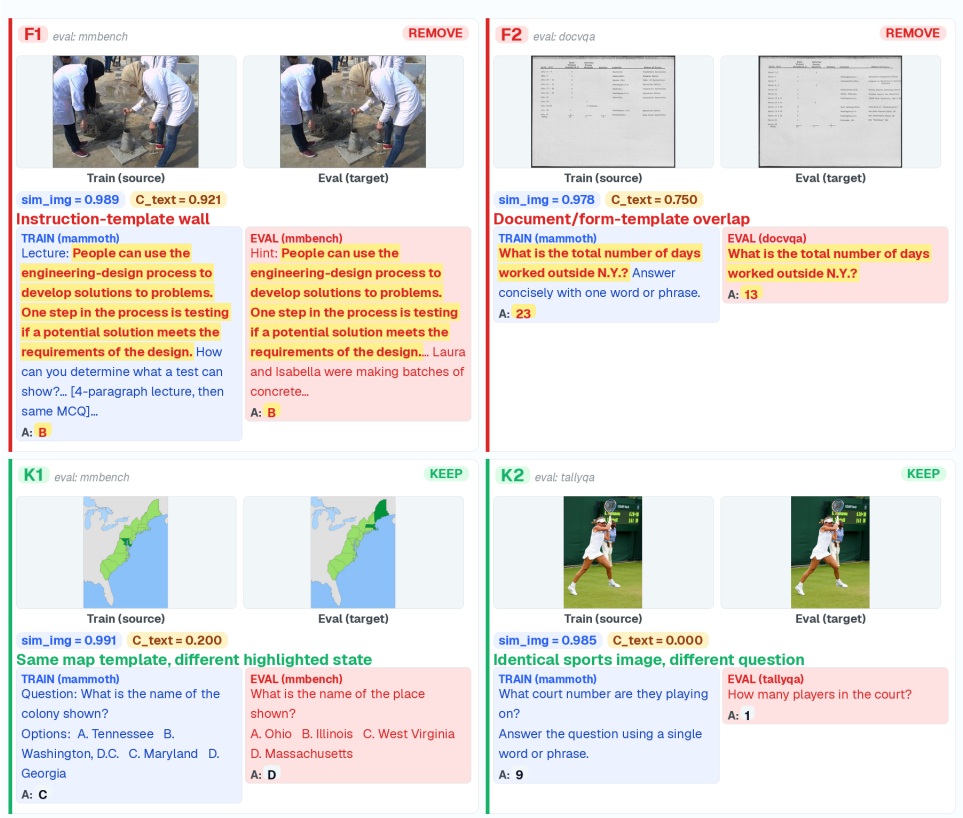


Figure 11 Representative flagged and non-flagged training/eval pairs across the joint-criterion regimes (Table 1) and the override patterns (fact-dump contamination, instruction-template wall, image-only RefCOCO-style). Each pair shows the eval sample alongside the matched training sample and the resulting sim_{img} and C_{text} scores.

against *every* eval sample, not just the image-nearest neighbour. A document is removed only if both gates pass; we bias toward recall, accepting false positives rather than missing a genuine leak.

Text containment is computed by concatenating question and answer, normalising (strip image tags and role markers, lowercase, collapse whitespace), and extracting word-level n-grams (n=4 by default; n=3 when the eval text has fewer than ten words). The score is directional,

$$C_{\text{text}}(x_e \to x_t) = \frac{|\mathcal{N}(x_e) \cap \mathcal{N}(x_t)|}{|\mathcal{N}(x_e)|}, \tag{A.2}$$

asking what fraction of the eval's n-grams appear in the training document. Directionality matters because training documents routinely embed eval Q/A verbatim inside longer chain-of-thought context, and the reverse direction would be diluted by that padding. Concatenating Q and A produces n-grams that span the Q/A boundary, a signature of the specific eval tuple rather than the template; question-only or answer-only matching collides with template-shaped fragments (*"how many bars are shown?"* matches every chart-counting sample). The Stage-2 broadcast across every eval sample (rather than just the Stage-1 nearest neighbour) matters when multiple eval samples share an image, as in AI2D, where the argmax can point to an eval whose text does not match while a different eval sharing the same image does.

We use τ_I=0.95 and τ_T=0.8 as global defaults, with per-benchmark overrides tuned by sweeping τ_T over $[0.4, 1.0]$ and inspecting the resulting precision profile via LLM-as-judge verification on a sampled set of flagged pairs. Three override patterns recur: fact-dump contamination and instruction-template walls (both requiring lowered τ_T tuned per benchmark), and benchmarks with no usable text signal (RefCOCO variants, PixMo, most scene/spatial evals), which rely on Stage 1 alone with a much higher $\tau_I \geq 0.995$.

Benchmark	% of corpus
vqa_v2	0.071%
mathvista	0.064%
ocr_vqa	0.025%
ocrbench_v2	0.011%
docvqa	0.008%
Other benchmarks (each <0.01%)	0.024%
Unique union (62 evaluations)	**0.19%**

Table 2 Per-benchmark share of the training corpus flagged by the decontamination pipeline. The final row reports the unique union across all 62 evaluations; per-row entries sum to slightly more than the union because a single training document may match multiple benchmarks.

Table 2 reports, for each benchmark, the share of training documents flagged and removed by the pipeline. Across the full training corpus, the pipeline removes 0.19% of documents as overlapping with one or more evaluation samples above the thresholds in Eq. A.1. The removed share is concentrated on benchmarks that share image sources with common training datasets (vqa_v2, mathvista, and ocr_vqa) and is negligible on benchmarks with distinct visual distributions.

B Training Hyperparameters

Table 3 reports the full set of hyperparameters used for every VLM pretraining run in this paper. The same recipe is applied to both the MAmmoTH-VL *baseline* and the DatologyAI-curated mixture, and to all three model scales (1B, 2B, 4B), unless a specific experiment (e.g. the context-length sweep in §5) explicitly varies a knob. Training is single-stage: the projector and language-model backbone are trained jointly from the start.

Architecture	
Vision encoder	SigLIP2-SO400M, 384×384, patch-14
LM backbone (1B / 2B / 4B)	Qwen3-0.6B / Qwen3-1.7B / Qwen3-4B
Projector	2-layer MLP
Visual tokenization	Dynamic tiling (aspect-ratio-aware), up to 12 tiles + thumbnail
Tokens per tile (pre / post pixel-shuffle)	729 / 243
Pixel-shuffle	1×3 (channels: 1,152 → 3,456)
Optimization	
Optimizer	AdamW
Peak learning rate (backbone / projector)	2×10^{-4} / 2×10^{-4}
LR schedule	WSD (warmup-stable-decay)
Warmup	50 steps (0.42% of total steps)
Weight decay	0.01
$\beta_1, \beta_2, \epsilon$	0.9, 0.999, 1×10^{-8}
Gradient clipping	1.0 global L2
Precision	bf16 autocast
Batching and schedule	
Token budget	25B tokens
Context length (default)	4,096
Context length (robustness sweep, §5)	8,192 / 16,384
Global batch size	512
Total training steps	12k
Sequence packing	Enabled
Data	
Baseline mixture	MAmmoTH-VL-12M single-image subset (10M samples)
Curated mixture	DatologyAI-curated on top of MAmmoTH-VL-12M single-image
Decontamination	Applied to all training data (see Appendix A)
Infrastructure	
Hardware	H100 GPUs
Parallelism	FSDP full-shard; activation checkpointing enabled; no tensor parallelism
Framework	PyTorch 2.9.1; Transformers 4.57.1; internal VLM training stack
Seeds per configuration	3

Table 3 Training hyperparameters. Values are shared across the 1B, 2B, and 4B runs and across *baseline* and *curated* mixtures, except where an experiment explicitly varies a single knob.

C Inference Settings and Full Results

This appendix documents (i) the inference configuration used for every evaluation number in the paper (§C.1), (ii) per-benchmark mean ± std scores for every internally trained model (1B, 2B, 4B at both MAmmoTH-VL *baseline* and DatologyAI-curated mixtures) and every external reference model on the full 20-eval public suite and the 9-capability DatBench suite (§C.2), (iii) per-benchmark mean response-token counts for the same model set, underlying the response-FLOPs axis of the inference-efficiency Pareto in §7 (§C.3), (iv) the FLOPs methodology and per-model accounting for both training-FLOPs and response-FLOPs (§C.4), (v) the context-length sweep underlying §5 (§C.5), and (vi) the case-study raw-number tables for grounding, counting, and BLINK referenced from §4 (§C.6). Internal checkpoints aggregate over three training seeds per configuration; external models are single runs as released. Everything reported in the main text is a summary of numbers in this appendix.

C.1 Inference settings

Table 4 reports the inference configuration used for our MAmmoTH-VL *baseline* and the DatologyAI-curated checkpoints. These runs use the same model-side inference recipe across benchmarks: zero-shot prompting, greedy decoding, and the same image preprocessing as training. Benchmark-specific prompt formatting, generation length caps, batch size, and scoring rules follow each evaluation's standard protocol. Public reference models (e.g. Qwen3-VL and InternVL-family models) are evaluated separately using their model-specific published decoding presets.

C.2 Per-benchmark scores across all models and scales

This subsection gives per-benchmark mean ± std scores for every internally trained model (MAmmoTH-VL *baseline* and DatologyAI-curated at 1B, 2B, 4B) alongside every external reference model, on (i) the full 20-eval public suite (Table 5) and (ii) the 9-capability DatBench suite (Table 6). The aggregated figures in the main text (§4, §5) take the Avg column or per-capability columns of these tables. The seed-variance aggregate quoted in §5 (mean per-capability seed std drops from 2.47 to 0.82, −67%, underlying Figure 8) is the mean of the per-capability Std cells of Table 6 at 2B.

C.2.1 All public benchmarks (20 evals)

Table 5 covers the 20-eval public suite shown in the hero Pareto (Figure 1); this is the union of the 11 IID benchmarks summarized in Figure 2 and the 9 OOD benchmarks summarized in Figure 3 (including BLINK and the ecommerce brand-ID component). Evals are rows, models are columns; MAmmoTH-VL/Datology pairs are the matched-compute *baseline*/ DatologyAI-curated runs at each scale.

C.2.2 DatBench (9 capabilities)

Table 6 gives the per-capability DatBench breakdown for every internal model and external reference. The 2B row pair underlies Figure 4; the 1B/2B/4B Datology rows underlie the scaling figure (Figure 6).

C.3 Mean response tokens per benchmark

Table 7 reports the mean number of generated response tokens per benchmark for every model in the comparison set, on the same 20-eval public suite as Table 5. The Avg column is the across-eval mean and is what drives the response-FLOPs axis in the inference-efficiency Pareto (Figure 10, §7): response-FLOPs $= 2 \cdot$ active params $\cdot$ Avg tokens.

Decoding	
Sampling strategy	Greedy
Temperature	0.0
Top-p	n/a
Top-k	n/a
Repetition penalty	none
Maximum new tokens	2,048
Stop sequences	none
Prompting	
System prompt	none
Prompt template	Benchmark-specific MCQ or generative template
Few-shot examples	0
Inference runtime	
Image preprocessing	Matches training
Precision	bf16 autocast
Batch size (inference)	64
Framework	`datvlmeval` native checkpoint inference
Decoding seed	n/a
Scoring	
Grounding (RefCOCO family)	center accuracy, IoU, Recall@$\{0.3, 0.5, 0.7\}$
Counting (CountBench)	exact / within-1 / within-3
Multi-image (BLINK)	per-category accuracy
Other DatBench axes	LM-as-judge

Table 4 Inference and evaluation settings for the DatologyAI-curated and MAmmoTH-VL *baseline* checkpoints. Public references are evaluated with their own model-specific decoding presets but the same benchmark-side evaluation protocols.

C.4 FLOPs methodology

We report two FLOPs metrics in the paper: *training FLOPs* on the hero Pareto's x-axis (Figure 1) and *response FLOPs* on the inference-efficiency Pareto's x-axis (Figure 10, §7). Both follow the standard transformer cost approximations:

$$F_{\text{train}} = 6 \cdot N \cdot D,$$
$$F_{\text{response}} = 2 \cdot N \cdot T,$$

where N is the model's active parameter count, D is its number of vision–language (VL) training tokens, and T is its mean generated-response token count averaged across the 20 evals in Table 7. The $6N$ approximation is the standard transformer-training cost (forward + backward + activation recompute); the $2N$ inference cost counts the decode-only forward pass per generated token.

We restrict D to *VL-stage* tokens so the comparison is apples-to-apples across models with different language-only pretraining histories: language-only LLM pretraining is folded into N (it is part of the backbone) and is not double-counted in D. Under this convention our matched-compute MAmmoTH-VL *baseline* and DatologyAI-curated runs both use $D = 25$B VL tokens at every scale.

All MAmmoTH-VL/Datology checkpoints share architecture: a Qwen3 LLM backbone (unfrozen during VL training), a 400M-parameter SigLIP-Large ViT, and a 2-layer MLP projector. Per-model N, D, T, and the

	Ours (mean ± std across seeds)						*External* (single run)						
Eval	MAmmoTH-VL 1B	Datology 1B	MAmmoTH-VL 2B	Datology 2B	MAmmoTH-VL 4B	Datology 4B	InternVL3-2B	InternVL3.5-2B	InternVL3.5-4B	Qwen3-VL-2B	Qwen3-VL-4B	Qwen3.5-2B	Qwen3.5-4B
RefCOCO	44.8 ± 0.7	82.9 ± 0.2	44.3 ± 11.5	85.1 ± 0.7	33.2 ± 10.4	89.2 ± 0.8	49.4	22.1	78.0	88.5	89.5	89.1	87.6
RefCOCOG	34.7 ± 4.4	77.0 ± 0.6	38.8 ± 11.3	79.5 ± 0.6	29.4 ± 9.5	83.3 ± 0.4	50.2	26.6	70.9	82.8	85.5	83.7	83.4
RefCOCO+	39.9 ± 0.7	74.5 ± 0.0	40.0 ± 10.8	76.2 ± 1.2	30.2 ± 10.7	82.7 ± 1.7	45.3	21.0	72.3	82.9	87.0	83.6	83.6
PixMo Points	2.4 ± 0.1	15.4 ± 0.6	1.5 ± 0.3	19.5 ± 1.2	2.8 ± 0.9	23.2 ± 1.9	3.6	10.9	13.4	9.0	9.8	10.2	14.8
MMBench	61.5 ± 0.5	63.3 ± 1.7	69.7 ± 1.2	70.3 ± 0.5	74.4 ± 1.2	76.7 ± 0.3	72.2	75.7	68.6	75.7	83.6	78.0	81.8
RealWorldQA	43.3 ± 1.5	53.7 ± 0.8	47.2 ± 1.6	57.9 ± 0.4	53.3 ± 0.7	61.8 ± 1.4	54.4	51.5	57.9	55.4	64.7	56.7	58.3
TextVQA	65.1 ± 0.5	67.7 ± 1.2	68.7 ± 0.3	73.4 ± 0.4	72.5 ± 1.7	75.0 ± 0.5	72.7	67.3	70.5	74.6	76.2	68.6	74.3
OCRBench	61.3 ± 1.0	65.8 ± 0.2	66.9 ± 0.7	70.6 ± 0.9	68.0 ± 1.4	73.7 ± 0.4	84.1	83.4	82.7	84.4	86.2	86.3	89.3
DocVQA	72.9 ± 1.2	80.8 ± 0.0	80.5 ± 0.4	85.1 ± 0.2	83.7 ± 0.4	88.9 ± 0.2	84.1	84.9	89.0	90.9	93.2	87.7	92.3
DetailCaps	63.2 ± 0.1	62.9 ± 0.1	63.1 ± 0.1	63.5 ± 0.3	62.8 ± 0.1	64.2 ± 0.0	64.2	64.1	63.8	63.9	63.0	61.8	57.2
CAPability	59.2 ± 0.2	61.0 ± 0.6	60.9 ± 0.1	61.7 ± 0.3	62.2 ± 0.4	64.5 ± 1.5	61.8	61.5	61.0	71.1	74.2	74.5	77.0
CVBench-2D	30.6 ± 2.1	59.8 ± 0.6	47.2 ± 2.4	65.2 ± 1.8	53.5 ± 1.9	71.6 ± 0.0	60.8	63.6	62.9	60.0	72.1	62.9	72.4
CVBench-3D	44.4 ± 11.2	75.1 ± 4.8	55.8 ± 3.1	76.3 ± 2.7	62.2 ± 4.8	78.3 ± 1.0	69.4	66.9	78.8	79.2	89.1	81.0	84.1
3DSRBench	32.5 ± 0.1	42.8 ± 0.8	32.9 ± 3.0	45.9 ± 0.3	42.0 ± 1.4	48.0 ± 0.0	38.8	43.0	32.8	43.2	46.5	47.0	45.0
CountBench	83.5 ± 0.2	88.8 ± 0.4	86.1 ± 0.4	89.9 ± 0.7	82.9 ± 1.0	90.2 ± 0.6	66.2	71.7	80.2	80.4	81.3	85.3	89.8
AI2D	61.7 ± 0.3	65.5 ± 0.3	70.5 ± 0.7	73.7 ± 0.2	74.5 ± 0.1	77.6 ± 0.4	75.6	74.7	78.7	74.2	81.1	77.4	76.5
ChartQA	72.4 ± 0.6	78.7 ± 0.2	78.9 ± 0.1	81.7 ± 0.1	81.4 ± 0.6	84.0 ± 0.5	78.4	74.8	84.0	76.8	82.8	83.1	82.3
MathVista	40.1 ± 0.0	45.6 ± 0.2	49.8 ± 0.8	53.6 ± 0.9	52.8 ± 2.6	57.4 ± 0.6	60.9	61.4	67.0	59.7	72.0	71.4	64.1
Brand ID	23.4 ± 7.8	34.0 ± 0.4	31.5 ± 4.0	36.7 ± 0.2	28.2 ± 3.3	37.1 ± 0.0	35.4	35.2	36.3	36.7	31.6	37.3	38.5
BLINK	40.3 ± 0.7	42.5 ± 0.3	41.7 ± 0.6	44.8 ± 0.7	44.3 ± 1.5	45.9 ± 0.7	44.2	52.6	57.8	55.7	66.6	62.0	63.5
Avg	**48.9** ± 0.7	**61.9** ± 0.3	**53.8** ± 1.4	**65.5** ± 0.0	**54.7** ± 1.4	**68.7** ± 0.3	**58.6**	**55.6**	**65.3**	**67.3**	**71.8**	**69.4**	**70.8**

Table 5 Per-benchmark scores on the full 20-eval public suite (the union of the 11 IID benchmarks in Figure 2 and the 9 OOD benchmarks in Figure 3; same set used by the hero Pareto in Figure 1). *MAmmoTH-VL* columns are the matched-compute MAmmoTH-VL-12M single-image *baseline*; *Datology* columns are the DatologyAI-curated mixture under the same recipe and compute. Internal cells show mean ± std across training seeds; external models are evaluated as released. The bottom **Avg** row reports each model's across-eval mean; the ± on Datology Avg cells is the std of the per-seed across-eval means (i.e., how much the suite-level average varies seed-to-seed), distinct from the mean per-cell std reported as the **DatBench** reliability headline in §5.

	Ours (mean ± std across seeds)						*External* (single run)						
Capability	MAmmoTH-VL 1B	Datology 1B	MAmmoTH-VL 2B	Datology 2B	MAmmoTH-VL 4B	Datology 4B	InternVL3-2B	InternVL3.5-2B	InternVL3.5-4B	Qwen3-VL-2B	Qwen3-VL-4B	Qwen3.5-2B	Qwen3.5-4B
Chart	32.6 ± 1.1	44.9 ± 0.4	39.8 ± 0.5	51.7 ± 0.9	45.2 ± 1.3	57.8 ± 0.1	49.1	55.6	62.9	56.5	64.6	65.8	69.0
Counting	77.8 ± 0.8	78.4 ± 0.5	80.0 ± 1.4	82.5 ± 0.8	81.6 ± 0.3	84.4 ± 0.9	82.9	62.8	82.6	84.3	89.1	83.5	80.3
Document	35.5 ± 0.4	44.2 ± 0.6	39.5 ± 0.9	49.1 ± 0.1	43.2 ± 0.7	52.2 ± 0.1	50.8	52.2	47.3	60.1	63.2	56.8	51.7
General	48.3 ± 0.4	51.1 ± 0.0	55.9 ± 0.5	57.5 ± 0.1	60.7 ± 0.4	62.8 ± 0.4	60.5	55.7	61.5	65.2	75.9	65.6	67.8
Grounding	31.0 ± 0.9	72.3 ± 1.1	18.3 ± 14.2	75.4 ± 1.7	29.1 ± 11.7	81.7 ± 0.9	51.0	30.6	79.0	82.6	87.5	81.2	84.2
Math	13.1 ± 0.5	14.1 ± 0.1	16.8 ± 0.2	19.1 ± 0.3	18.2 ± 0.8	23.1 ± 0.3	16.9	24.9	25.0	22.8	33.3	33.6	24.7
Scene	57.3 ± 1.0	61.4 ± 0.4	61.1 ± 0.2	67.8 ± 0.5	63.5 ± 2.2	72.0 ± 0.5	67.4	60.5	64.9	79.9	83.2	58.1	40.0
Spatial	36.2 ± 2.9	42.5 ± 0.7	41.0 ± 2.6	45.3 ± 2.6	48.3 ± 1.2	55.2 ± 1.1	38.6	30.1	54.0	37.7	39.1	31.1	29.1
Table	35.5 ± 1.0	36.8 ± 0.0	36.2 ± 1.8	42.4 ± 0.5	44.1 ± 1.6	49.7 ± 0.4	37.6	51.1	57.9	49.5	63.3	50.9	24.5
Avg	**40.8** ± 0.5	**49.5** ± 0.3	**43.2** ± 1.7	**54.5** ± 0.5	**48.2** ± 1.4	**59.9** ± 0.0	**50.5**	**47.1**	**59.4**	**59.8**	**66.6**	**58.5**	**52.4**

Table 6 Per-capability scores on the 9-capability **DatBench** suite. *MAmmoTH-VL* columns are the matched-compute MAmmoTH-VL-12M single-image *baseline*; *Datology* columns are the DatologyAI-curated mixture. Internal cells show mean ± std across training seeds; external models are evaluated as released. The 2B MAmmoTH-VL/Datology column pair underlies the per-capability bars in Figure 4; the 1B/2B/4B Datology columns underlie the scaling story in Figure 6. Capability column abbreviations: Chart = Chart Understanding, Document = Document Understanding, Scene = Scene OCR, Table = Diagrams & Tables, Math = Math & Logic, Spatial = Spatial Reasoning. The Avg column ± on Datology rows is the std of per-seed across-capability means; the mean per-capability std (the −67% headline in §5) is the row-wise mean of the Std cells.

resulting F_{train} and F_{response} are summarized in Table 8; sources for N and D on external references are model cards / release notes.

As a worked example for the *curated* 2B: $N = 2.10 \times 10^9$ (1.7B Qwen3 LLM + 0.4B SigLIP ViT), $D = 25 \times 10^9$ VL tokens, $T = 42.1$ (the 20-eval mean from Table 7), giving $F_{\text{train}} = 6 \cdot 2.10 \times 10^9 \cdot 25 \times 10^9 = 3.15 \times 10^{20}$ FLOPs and $F_{\text{response}} = 2 \cdot 2.10 \times 10^9 \cdot 42.1 = 1.77 \times 10^{11}$ FLOPs.

C.5 Context-length sweep

Tables 9 and 10 give the per-benchmark / per-capability breakdown of the 4k / 8k / 16k context-length sweep at 2B, for both MAmmoTH-VL *baseline* and DatologyAI-curated mixtures. The Avg row aggregates underlie the curation-gain figure (Figure 7) and the seed-variance-vs-context-length figure (Figure 8) in §5.

C.6 Case-study and OOD raw numbers

This subsection lifts out the per-seed and per-metric tables for grounding (§4), counting, and BLINK (§4 OOD subsection), which sit in the appendix to keep the main text readable. Counting is included here in full (no main-text subsection); the grounding deep-dive in §4 cross-references this appendix as a parallel instance of the same coherent-across-metrics pattern.

	Ours						*External*						
Eval	MAmmoTH-VL 1B	Datology 1B	MAmmoTH-VL 2B	Datology 2B	MAmmoTH-VL 4B	Datology 4B	InternVL3-2B	InternVL3.5-2B	InternVL3.5-4B	Qwen3-VL-2B	Qwen3-VL-4B	Qwen3.5-2B	Qwen3.5-4B
RefCOCO	24.1	19.6	21.5	22.1	21.2	22.5	29.5	150.6	23.1	22.5	22.2	23.7	744.3
RefCOCOG	23.6	19.6	21.2	21.9	21.7	22.6	29.9	120.7	23.1	22.5	22.3	23.8	830.8
RefCOCO+	24.8	19.6	21.7	22.1	21.5	22.5	32.3	147.5	23.3	22.6	22.1	23.7	783.5
PixMo Points	9.8	11.7	9.1	11.7	29.3	11.5	35.0	132.1	34.5	14.7	44.3	79.1	1129.5
MMBench	1.1	1.0	1.0	1.0	1.2	1.0	11.6	132.6	108.5	100.2	155.6	171.3	2684.6
RealWorldQA	1.2	1.1	1.2	1.1	1.2	1.3	5.3	70.6	30.2	6.2	57.4	86.3	2167.0
TextVQA	7.2	6.0	5.3	4.1	8.3	4.9	3.2	22.3	42.4	6.7	5.0	35.8	702.0
OCRBench	6.5	9.9	7.1	8.5	7.0	12.4	9.9	25.5	28.5	37.0	34.7	36.5	549.4
DocVQA	5.6	5.9	5.2	5.4	5.5	5.8	6.1	35.5	6.3	6.5	5.7	38.5	360.1
DetailCaps	456.3	138.3	474.9	298.9	468.7	325.4	173.7	178.8	155.1	282.6	358.0	408.6	1229.0
CAPability	517.6	396.7	526.7	425.5	520.8	202.2	139.7	210.8	159.1	289.8	373.7	435.7	1271.3
CVBench-2D	1.5	1.0	1.0	1.0	1.1	1.0	3.6	101.1	28.5	9.1	103.4	123.1	2832.5
CVBench-3D	1.4	1.0	1.0	1.0	1.0	1.0	3.1	188.8	113.5	5.0	11.9	103.8	1564.0
3DSRBench	1.1	1.0	1.0	1.0	20.0	1.0	4.1	147.2	51.4	19.5	40.1	179.6	2921.6
CountBench	1.8	1.2	1.2	1.4	1.5	1.4	2.8	92.2	86.5	50.8	8.7	42.5	562.8
AI2D	1.0	1.0	1.0	1.0	1.0	1.0	6.0	157.9	229.5	302.3	88.3	338.0	1216.1
ChartQA	4.0	3.5	3.5	4.3	3.5	3.6	3.7	18.8	18.4	7.6	11.8	78.1	658.3
MathVista	14.5	2.8	5.8	2.6	37.7	2.1	77.5	278.9	255.0	446.0	458.0	524.4	1282.9
Brand ID	11.2	6.2	5.7	5.4	16.1	5.9	14.1	114.0	76.1	84.8	68.5	162.8	778.0
BLINK	31.1	1.0	1.5	1.0	3.5	1.1	8.6	244.8	214.6	295.0	285.3	472.2	1410.8
Avg	**57.3**	**32.4**	**55.8**	**42.1**	**59.6**	**32.5**	**30.0**	**128.5**	**85.4**	**101.6**	**108.9**	**169.4**	**1283.9**

Table 7 Mean generated-response token count per benchmark on the same 20-eval public suite as Table 5, averaged across seeds for our checkpoints and across released runs for externals. The bottom **Avg** row is the across-eval mean used as the response-tokens factor in the response-FLOPs proxy ($2 \cdot$ active params $\cdot$ Avg tokens) for the inference-efficiency Pareto in Figure 10 (§7). The two captioning evals (DetailCaps, CAPability) and the post-trained references (notably Qwen3.5-4B) account for most of the variation in the Avg row.

Model	LLM backbone	N	D (VL tokens)	T (resp tokens)	F_{train}	F_{response}
MAmmoTH-VL *Baseline* 1B	Qwen3-0.6B	1.00B	25B	57.3	1.50×10^{20}	1.15×10^{11}
Datology Curation 1B	Qwen3-0.6B	1.00B	25B	32.4	1.50×10^{20}	6.48×10^{10}
MAmmoTH-VL *Baseline* 2B	Qwen3-1.7B	2.10B	25B	55.8	3.15×10^{20}	2.34×10^{11}
Datology Curation 2B	Qwen3-1.7B	2.10B	25B	42.1	3.15×10^{20}	1.77×10^{11}
MAmmoTH-VL *Baseline* 4B	Qwen3-4B	4.40B	25B	59.6	6.60×10^{20}	5.24×10^{11}
Datology Curation 4B	Qwen3-4B	4.40B	25B	32.5	6.60×10^{20}	2.86×10^{11}
InternVL3-2B (MPO)	Qwen2.5-1.5B	2.09B	244B	30.0	3.06×10^{21}	1.25×10^{11}
InternVL3.5-2B (CascadeRL)	Qwen3-1.7B	2.30B	381B	128.5	5.26×10^{21}	5.91×10^{11}
InternVL3.5-4B (CascadeRL)	Qwen3-4B	4.30B	381B	85.4	9.83×10^{21}	7.34×10^{11}
Qwen3-VL-2B	Qwen3-1.7B	2.10B	2.17T	101.6	2.73×10^{22}	4.27×10^{11}
Qwen3-VL-4B	Qwen3-4B	4.30B	2.17T	108.9	5.60×10^{22}	9.36×10^{11}
Qwen3.5-2B	Qwen3-1.7B	2.00B	4.00T	169.4	4.80×10^{22}	6.77×10^{11}
Qwen3.5-4B	Qwen3-4B	4.00B	4.00T	1283.9	9.60×10^{22}	1.03×10^{13}

Table 8 Per-model active parameters (N), VL training tokens (D), 20-eval mean response-token count (T, from Table 7), and the resulting training FLOPs $F_{\text{train}} = 6ND$ and response FLOPs $F_{\text{response}} = 2NT$ used in Figure 1 and Figure 10 respectively. Sources for external N and D are the corresponding model cards / release notes.

C.6.1 Grounding (RefCOCO) per-metric

C.6.2 Counting (CountBench) per-seed

CountBench scores at exact match, within-1, and within-3 tolerances; jointly they distinguish a narrow gain (one tolerance only) from a broad one. Curation lifts every tolerance: exact match from $86.08 \pm 0.51\%$ to $89.88 \pm 0.92\%$ (+3.8pp), within-1 from $95.32 \pm 0.89\%$ to $97.22 \pm 0.51\%$ (+1.9pp), and within-3 from $97.90 \pm 0.42\%$ to $99.12 \pm 0.12\%$ (+1.2pp). Every *curated* seed exceeds every *baseline* seed at all three tolerances. The gain shows up at the strictest threshold and carries through the looser ones, the same pattern observed in the grounding deep-dive (§4).

C.6.3 BLINK per-seed

	4k context		8k context		16k context	
Eval	MAmmoTH-VL 2B	Datology 2B	MAmmoTH-VL 2B	Datology 2B	MAmmoTH-VL 2B	Datology 2B
RefCOCO	44.3 ± 11.5	85.1 ± 0.7	29.4 ± 18.5	81.9 ± 1.9	13.3 ± 13.9	81.2 ± 2.5
RefCOCOG	38.8 ± 11.3	79.5 ± 0.6	28.9 ± 15.8	75.2 ± 0.0	14.3 ± 15.2	74.4 ± 2.3
RefCOCO+	40.0 ± 10.8	76.2 ± 1.2	26.3 ± 15.4	73.6 ± 1.4	12.3 ± 13.2	71.8 ± 2.5
PixMo Points	1.5 ± 0.3	19.5 ± 1.2	0.9 ± 0.9	2.3 ± 1.2	0.4 ± 0.0	1.4 ± 0.3
MMBench	69.7 ± 1.2	70.3 ± 0.5	69.1 ± 0.5	69.9 ± 0.8	68.8 ± 0.2	69.3 ± 0.2
RealWorldQA	47.2 ± 1.6	57.9 ± 0.4	46.7 ± 1.1	54.5 ± 0.0	48.4 ± 1.6	51.7 ± 1.6
TextVQA	68.7 ± 0.3	73.4 ± 0.4	69.5 ± 1.3	70.0 ± 0.5	69.0 ± 1.1	68.7 ± 1.4
OCRBench	66.9 ± 0.7	70.6 ± 0.9	64.8 ± 1.9	67.6 ± 0.6	63.0 ± 0.1	64.3 ± 0.4
DocVQA	80.5 ± 0.4	85.1 ± 0.2	79.0 ± 0.8	81.0 ± 0.6	77.4 ± 0.8	80.4 ± 0.5
DetailCaps	63.1 ± 0.1	63.5 ± 0.3	62.8 ± 0.1	62.9 ± 0.2	62.8 ± 0.1	63.4 ± 0.0
CAPability	60.9 ± 0.1	61.7 ± 0.3	60.4 ± 0.3	60.5 ± 0.5	60.0 ± 0.3	61.6 ± 0.6
CVBench-2D	47.2 ± 2.4	65.2 ± 1.8	39.7 ± 7.2	61.3 ± 1.6	40.1 ± 6.2	56.2 ± 1.1
CVBench-3D	55.8 ± 3.1	76.3 ± 2.7	54.8 ± 2.4	60.8 ± 9.7	46.1 ± 3.6	58.9 ± 0.2
3DSRBench	32.9 ± 3.0	45.9 ± 0.3	31.4 ± 2.1	42.8 ± 0.0	29.6 ± 4.4	39.5 ± 0.5
CountBench	86.1 ± 0.4	89.9 ± 0.7	85.9 ± 0.6	85.8 ± 0.1	83.9 ± 0.3	85.4 ± 0.1
AI2D	70.5 ± 0.7	73.7 ± 0.2	69.7 ± 0.1	70.6 ± 0.0	68.2 ± 0.5	68.9 ± 0.2
ChartQA	78.9 ± 0.1	81.7 ± 0.1	78.4 ± 0.4	77.2 ± 0.5	77.2 ± 0.7	77.4 ± 0.0
MathVista	49.8 ± 0.8	53.6 ± 0.9	49.6 ± 2.3	47.1 ± 1.4	46.6 ± 0.7	46.4 ± 1.6
Brand ID	31.5 ± 4.0	36.7 ± 0.2	27.2 ± 5.6	33.7 ± 1.9	30.8 ± 5.5	28.6 ± 1.6
BLINK	41.7 ± 0.6	44.8 ± 0.7	41.7 ± 0.1	43.3 ± 0.1	42.2 ± 1.2	43.4 ± 0.5
Avg	**53.8** ± 1.4	**65.5** ± 0.0	**50.8** ± 2.4	**61.1** ± 0.8	**47.7** ± 2.8	**59.6** ± 0.3

Table 9 Per-benchmark scores on the 20-eval public suite at 2B across the 4k / 8k / 16k context-length sweep, *baseline* (MAmmoTH-VL) vs. DatologyAI-curated. Cells are mean ± std across training seeds (Datology has 3 / 2 / 2 seeds at 4k / 8k / 16k respectively). The Avg row gives a +11.7 / +10.3 / +11.9pp curation gain at 4k / 8k / 16k. Underlies Figure 7 (suite-level deltas) and Figure 8 (cross-seed std).

	4k context		8k context		16k context	
Capability	MAmmoTH-VL 2B	Datology 2B	MAmmoTH-VL 2B	Datology 2B	MAmmoTH-VL 2B	Datology 2B
Chart	39.8 ± 0.5	51.7 ± 0.9	37.8 ± 0.7	38.4 ± 0.3	36.5 ± 1.8	39.9 ± 0.5
Counting	80.0 ± 1.4	82.5 ± 0.8	77.3 ± 1.8	82.3 ± 0.4	76.7 ± 0.5	81.6 ± 0.2
Document	39.5 ± 0.9	49.1 ± 0.1	37.5 ± 1.6	41.4 ± 0.5	37.9 ± 0.7	39.8 ± 0.2
General	55.9 ± 0.5	57.5 ± 0.1	55.5 ± 0.3	56.3 ± 0.7	52.8 ± 0.4	54.5 ± 0.9
Grounding	18.3 ± 14.2	75.4 ± 1.7	19.9 ± 12.6	72.3 ± 0.9	8.5 ± 9.2	70.4 ± 0.4
Math	16.8 ± 0.2	19.1 ± 0.3	16.5 ± 0.8	15.8 ± 0.2	16.3 ± 0.9	17.1 ± 0.9
Scene	61.1 ± 0.2	67.8 ± 0.5	57.0 ± 3.1	62.1 ± 0.5	55.5 ± 2.0	61.7 ± 1.6
Spatial	41.0 ± 2.6	45.3 ± 2.6	36.2 ± 5.2	42.7 ± 2.0	31.4 ± 7.4	43.2 ± 2.3
Table	36.2 ± 1.8	42.4 ± 0.5	35.3 ± 6.0	38.9 ± 0.5	33.4 ± 0.9	33.5 ± 1.8
Avg	**43.2** ± 1.7	**54.5** ± 0.5	**41.4** ± 2.3	**50.0** ± 0.5	**38.8** ± 1.8	**49.1** ± 0.6

Table 10 Per-capability scores on the 9-capability **DatBench** suite at 2B across the 4k / 8k / 16k context-length sweep. Cells are mean ± std across training seeds (Datology has 3 / 2 / 2 seeds at 4k / 8k / 16k respectively). The Avg row gives a +11.3 / +8.6 / +10.3pp curation gain at 4k / 8k / 16k. Underlies Figure 7 and Figure 8.

Metric	MAmmoTH-VL Mean ± Std	Datology Mean ± Std
`recall@0.3`	57.61 ± 16.99	85.92 ± 0.68
`recall@0.5`	41.04 ± 13.69	80.29 ± 1.01
`recall@0.7`	14.55 ± 4.85	69.96 ± 0.88
`center_acc`	69.72 ± 12.25	91.03 ± 0.40

Table 11 Grounding metrics on the RefCOCO subset, *baseline* vs. *curated*, in percentage units. Aggregated across the three RefCOCO splits used in §4.

	Exact	Within-1	Within-3
MAmmoTH-VL ts0	85.54	95.93	98.37
MAmmoTH-VL ts1	86.15	94.30	97.76
MAmmoTH-VL ts2	86.56	95.72	97.56
MAmmoTH-VL (Mean ± Std)	86.08 ± 0.51	95.32 ± 0.89	97.90 ± 0.42
Datology ts0	89.00	96.74	98.98
Datology ts1	89.82	97.15	99.19
Datology ts2	90.84	97.76	99.19
Datology (Mean ± Std)	89.88 ± 0.92	97.22 ± 0.51	99.12 ± 0.12
Abs. Gain	+3.80	+1.90	+1.22

Table 12 CountBench per-seed results across exact, within-1, and within-3 tolerances; all values in percentage units.

Training Mix	ts0	ts1	ts2	Mean	Std	Abs. Gain vs MAmmoTH-VL
MAmmoTH-VL	41.29	41.35	42.35	41.66	0.59	+0.00
Datology	44.03	44.87	45.34	44.75	0.67	+3.09

Table 13 BLINK per-seed accuracy in percentage units. Aggregate DatBench-style accuracy across all BLINK categories; the per-category breakdown referenced in §4 is summarized inline rather than tabulated here.

D Limitations and Future Work

The results in this paper are at 1B, 2B, and 4B parameter scales, on a single base corpus (MAmmoTH-VL-12M single-image subset), one backbone family (Qwen3 LM with SigLIP2 vision encoder), and one training recipe. The robustness evidence in §5 indicates the qualitative picture carries. A direct demonstration across backbones and on additional open VLM corpora is future work. The pipeline does not extend to multi-image or interleaved image-text training. The BLINK transfer in §4 indicates curation applied to those data types would further improve performance.