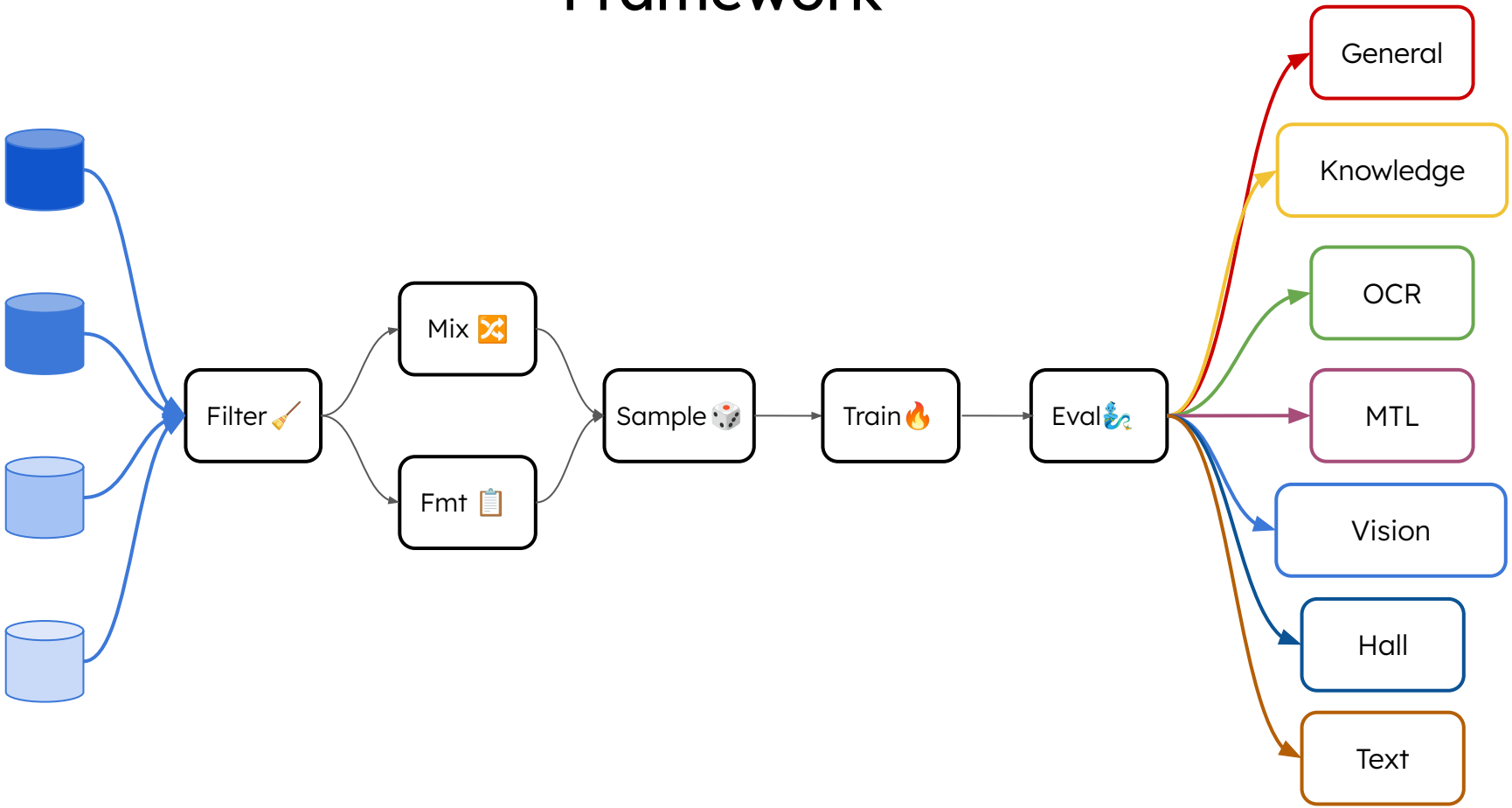
DataComp-VLM:
Improved Open Datasets for Vision-Language Models

TLDR.
Benchmarking curation strategies for building the best training dataset for single-stage pretraining of VLMs!

Framework


Filter 🧹
Mix 🔀
Fmt 📋
Sample 🎲
Train 🔥
Eval 🐲
General
Knowledge
OCR
MTL
Vision
Hall
Text

Data Sources

Multimodal Documents



The match between Tottenham Spurs vs Chelsea will kick off from 16:30 at Tottenham Hotspur Stadium, London.



The derby had been played 54 times and the Blues have dominated the Spurs. Out of 54 matches played, Chelsea has won 28 times and Spurs had only won 7 times. The remaining 19 matches had ended in draw.

However, in recent 5 meetings, Spurs had won 3 times where Chelsea had won the other two times. …

Image-Caption Pairs



cats with different colored through golden this ca...



cat, white, and eyes image



Coby the Cat with Piercing Blue Eyes, over 1 Milli...



heterochromia-cat-cross-eyed-alos-17

Text-only Data

FineWeb

The finest collection of data the web has to offer



Multimodal Instruction-Tuning

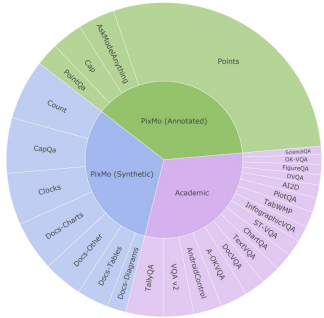


Figure 4. Datasets used for fine-tuning, shown in proportion to their sampling rates. Green denotes human-annotated data we collected, blue denotes synthetic data we generated, and purple represents pre-existing academic datasets. PixMo-Docs has been subdivided into charts, tables, diagrams, and other.

Data Sources

Multimodal Documents



The match between Tottenham Spurs vs Chelsea will kick off from 16:30 at Tottenham Hotspur Stadium, London.



The derby had been played 54 times and the Blues have dominated the Spurs. Out of 54 matches played, Chelsea has won 28 times and Spurs had only won 7 times. The remaining 19 matches had ended in draw.

However, in recent 5 meetings, Spurs had won 3 times where Chelsea had won the other two times. …

Image-Caption Pairs



cats with different colored through golden this ca...



cat, white, and eyes image



Coby the Cat with Piercing Blue Eyes, over 1 Milli...



heterochromia-cat-cross-eyed-alos-17

Text-only Data

FineWeb

The finest collection of data the web has to offer



Multimodal Instruction-Tuning

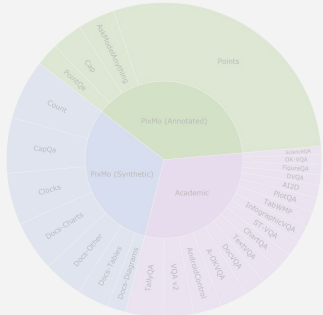


Figure 4. Datasets used for fine-tuning, shown in proportion to their sampling rates. Green denotes human-annotated data we collected, blue denotes synthetic data we generated, and purple represents pre-existing academic datasets. PixMo-Docs has been subdivided into charts, tables, diagrams, and other.

DataComp-CLIP

Data Sources

Multimodal Documents



The match between Tottenham Spurs vs Chelsea will kick off from 16:30 at Tottenham Hotspur Stadium, London.



The derby had been played 54 times and the Blues have dominated the Spurs. Out of 54 matches played, Chelsea has won 28 times and Spurs had only won 7 times. The remaining 19 matches had ended in draw.

However, in recent 5 meetings, Spurs had won 3 times where Chelsea had won the other two times. …

Image-Caption Pairs



cats with different colored through golden this ca...



cat, white, and eyes image



Coby the Cat with Piercing Blue Eyes, over 1 Milli...



heterochromia-cat-cross-eyed-alos-17

Text-only Data

FineWeb

The finest collection of data the web has to offer

DataComp-LM

Multimodal Instruction-Tuning

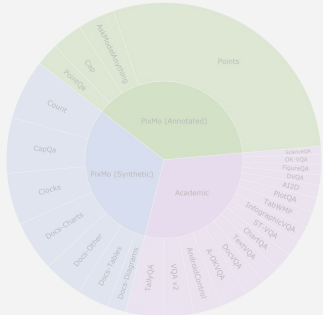


Figure 4. Datasets used for fine-tuning, shown in proportion to their sampling rates. Green denotes human-annotated data we collected, blue denotes synthetic data we generated, and purple represents pre-existing academic datasets. PixMo-Docs has been subdivided into charts, tables, diagrams, and other.

Data Sources

Multimodal Documents



The match between Tottenham Spurs vs Chelsea will kick off from 16:30 at Tottenham Hotspur Stadium, London.



The derby had been played 54 times and the Blues have dominated the Spurs. Out of 54 matches played, Chelsea has won 28 times and Spurs had only won 7 times. The remaining 19 matches had ended in draw.

However, in recent 5 meetings, Spurs had won 3 times where Chelsea had won the other two times. …

Image-Caption Pairs



cats with different colored through golden this ca...



cat, white, and eyes image



Coby the Cat with Piercing Blue Eyes, over 1 Milli...



heterochromia-cat-cross-eyed-alos-17

Text-only Data

FineWeb

The finest collection of data the web has to offer

Multimodal Instruction-Tuning



Figure 4. Datasets used for fine-tuning, shown in proportion to their sampling rates. Green denotes human-annotated data we collected, blue denotes synthetic data we generated, and purple represents pre-existing academic datasets. PixMo-Docs has been subdivided into charts, tables, diagrams, and other.

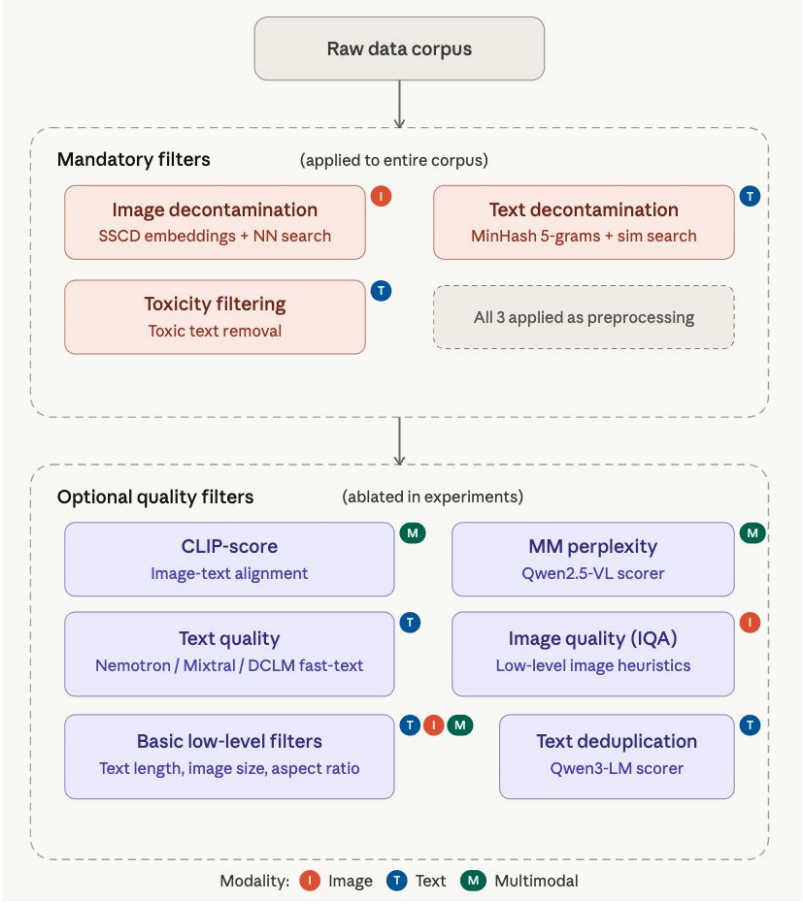
DataComp-VLM

Benchmark Scales

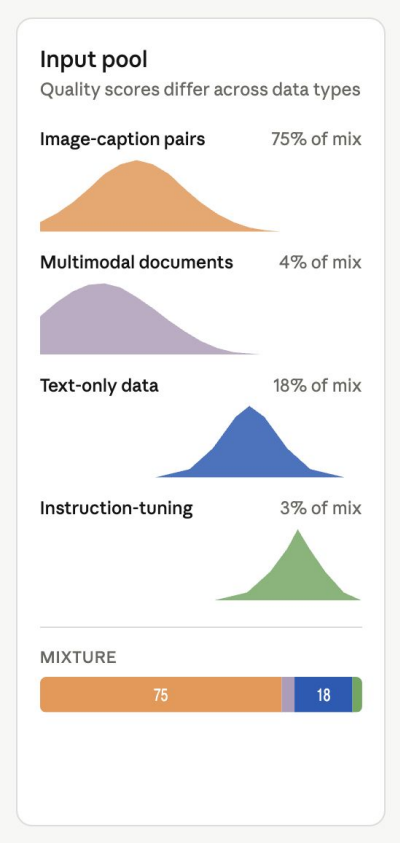
SCALING LADDER — 4 COMPUTE SCALES

Small		Medium		Large		X-Large (Hero)	
Pool	187.5B tok	Pool	750B tok	Pool	3T tok	Pool	~6T tok
Samples	125M	Samples	500M	Samples	2B	Samples	~4B
Token budget	6.25B	Token budget	25B	Token budget	100B	Token budget	200B
Model	1B	Model	2B	Model	4B	Model	8B
A100 hours	320	A100 hours	1,280	A100 hours	5,120	A100 hours	10,240

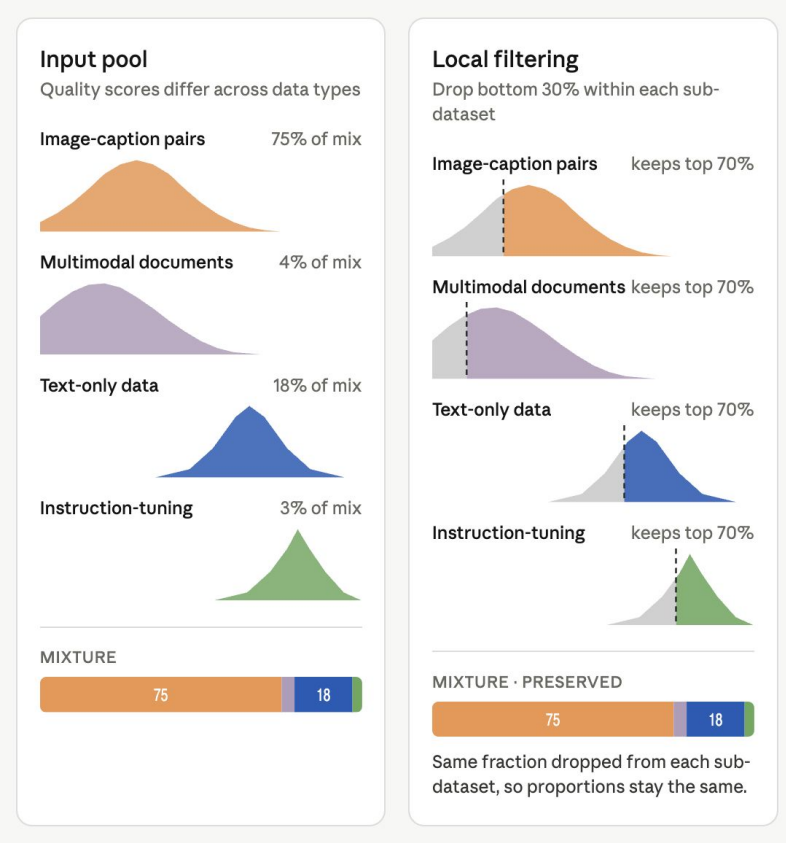
Filtering


Raw data corpus
Mandatory filters
(applied to entire corpus)
Image decontamination
SSCD embeddings + NN search
I
Text decontamination
MinHash 5-grams + sim search
T
Toxicity filtering
Toxic text removal
T
All 3 applied as preprocessing
Optional quality filters
(ablated in experiments)
CLIP-score
Image-text alignment
M
MM perplexity
Qwen2.5-VL scorer
M
Text quality
Nemotron / Mixtral / DCLM fast-text
T
Image quality (IQA)
Low-level image heuristics
I
Basic low-level filters
Text length, image size, aspect ratio
T I M
Text deduplication
Qwen3-LM scorer
T
Modality: I Image T Text M Multimodal

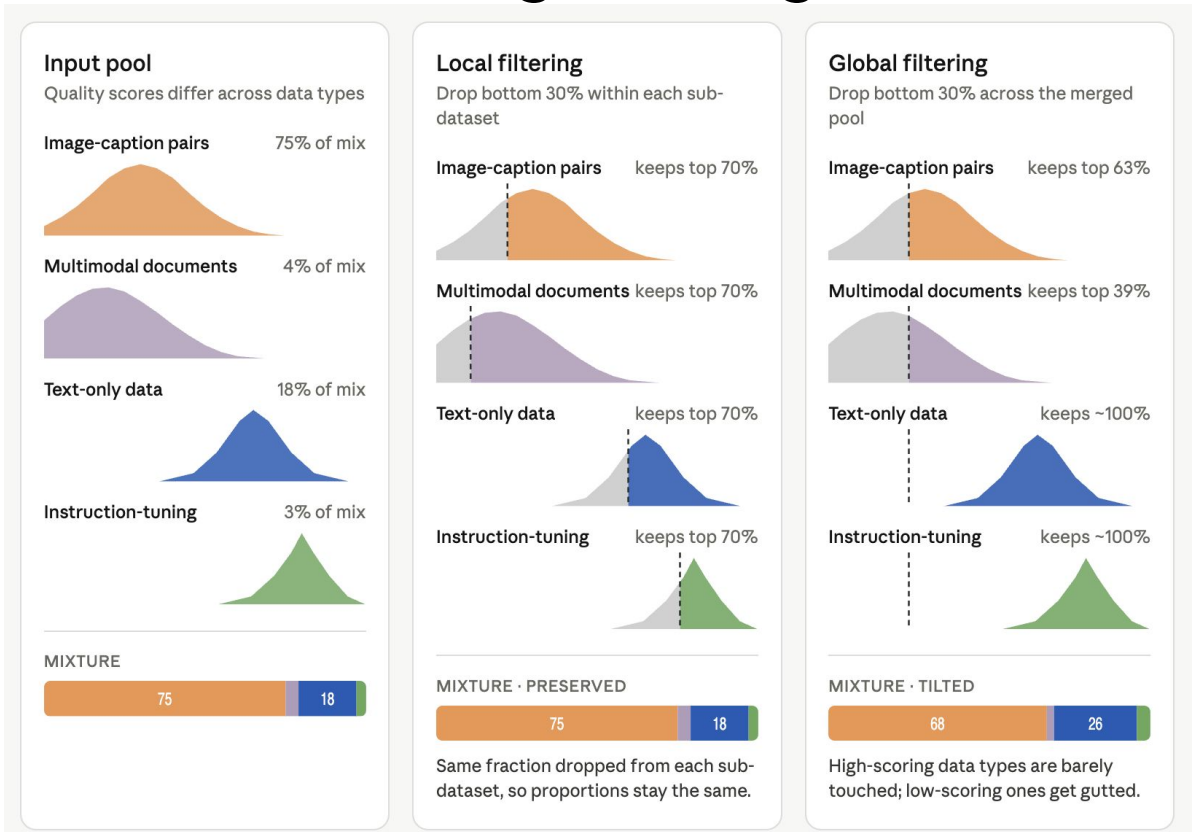
Filtering Strategies



Filtering Strategies

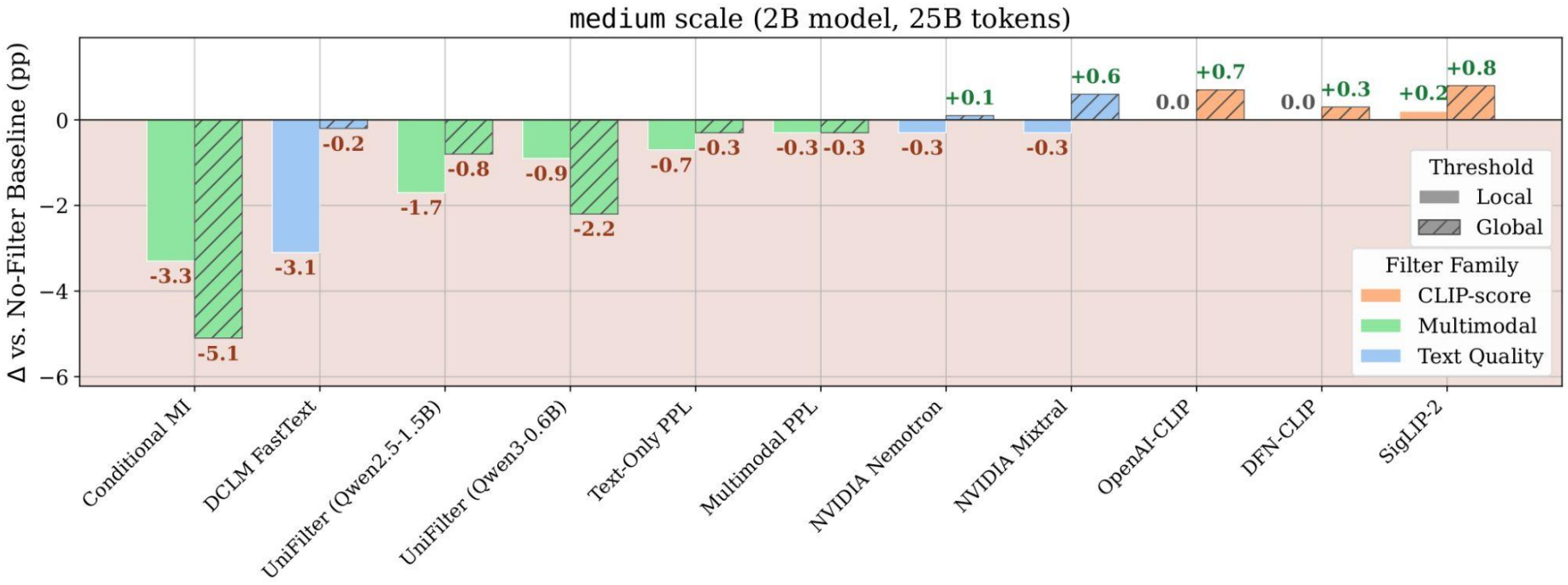


Filtering Strategies



Both methods retain 70% of samples. The same percentile cut maps to **four absolute thresholds** in the local case, and to **one absolute threshold** in the global case — implicitly inducing a new mixture.

Filtering is fragile!



Mixing

STEP 1 — RAW CORPUS COMPOSITION (LENGTH-PROPORTIONAL SAMPLING)



Image-caption pairs (75%) · MM documents (4%) · Text-only (18%) · Instruction-tuning (3%)

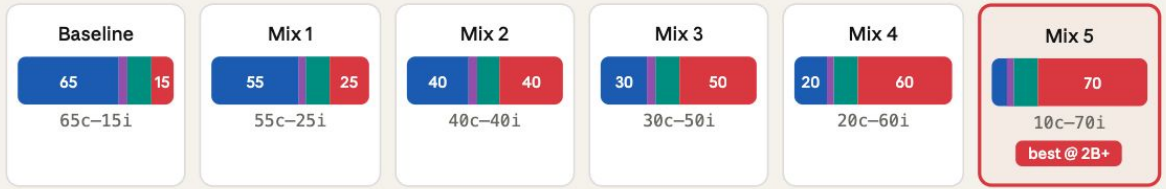
Heavily dominated by image-caption pairs from web crawls (DataComp, ReLAION). Instruction-tuning data nearly absent.

↓

STEP 2 — DATA-TYPE LINE SEARCH: IC% ↔ IT%

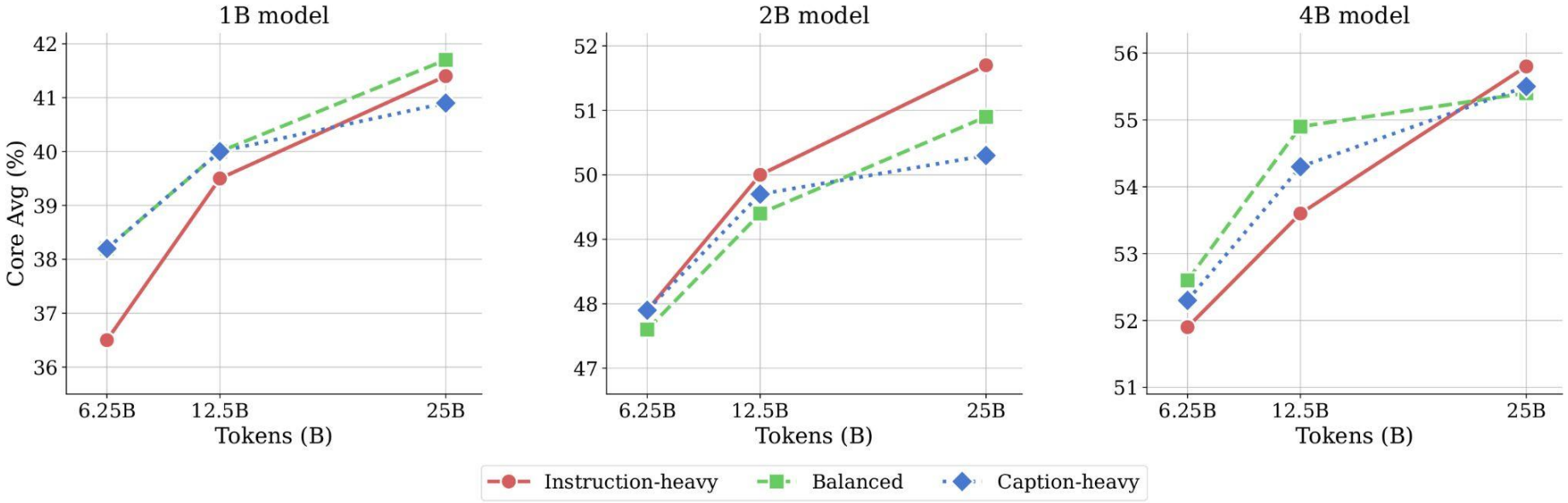
Image-caption ---------- *gradually trade* ---------- Instruction-tuning

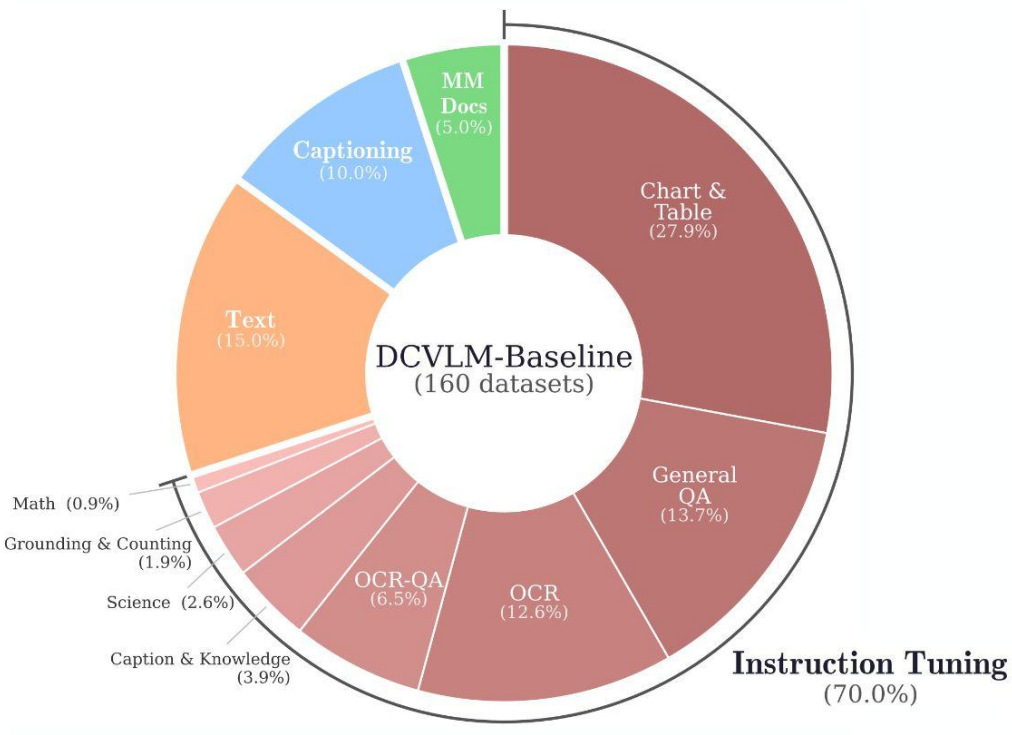
MMDoc fixed at 5% fixed · Text fixed at 15% fixed

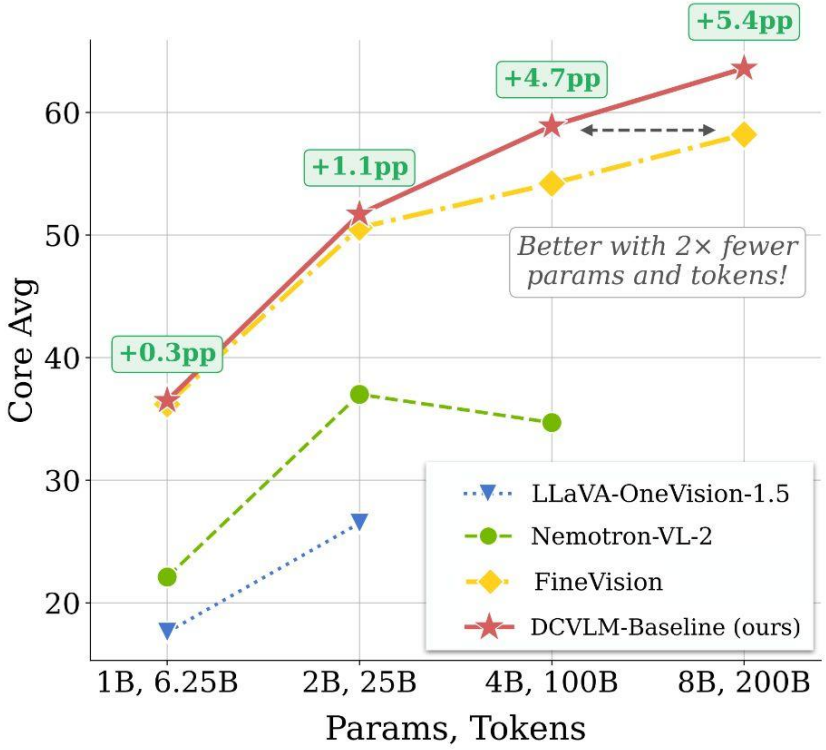


9 mixtures evaluated across **3 model sizes** (1B / 2B / 4B) × **3 token budgets** (6.25B / 12.5B / 25B). IT-heavy mixes scale better — at 2B+ parameters, 10c–70i consistently dominates.

Mixing gives the biggest gains!




MM Docs (5.0%)
Captioning (10.0%)
Chart & Table (27.9%)
Text (15.0%)
DCVLM-Baseline
(160 datasets)
General QA (13.7%)
OCR-QA (6.5%)
OCR (12.6%)
Math (0.9%)
Grounding & Counting (1.9%)
Science (2.6%)
Caption & Knowledge (3.9%)
Instruction Tuning
(70.0%)


+5.4pp
+4.7pp
+1.1pp
+0.3pp
Better with 2× fewer params and tokens!
Core Avg
60
50
40
30
20
1B, 6.25B
2B, 25B
4B, 100B
8B, 200B
Params, Tokens
LLaVA-OneVision-1.5
Nemotron-VL-2
FineVision
DCVLM-Baseline (ours)

Thanks for listening!

Hope to see you in Sydney 🇦🇺