MONET

A Massive, Open, Non-redundant and Enriched Text-to-image dataset

NeurIPS 2026 · Evaluations & Datasets

Benjamin Aubin, Gonzalo Iñaki Quintana, Onur Tasar, Sanjeev Sreetharan, Urszula Czerwinska, Damien Henry, Clément Chadebec

Jasper Research

Why MONET?

jasper

The gap

- Training text-to-image (T2I) models requires large curated datasets (strong visual diversity, high-quality images, detailed captions).
- Collecting, filtering, deduplicating, re-captioning at scale is expensive and time consuming → hinders open, accessible, transparent and reproducible research
- LAION, COYO: billions of pairs, but uncurated, redundant, with short noisy alt-text
- ShareGPT4V: rich captions, but only 1.2M images and a single captioner that biases prompts
- No open dataset that is filtered, deduplicated and multi-VLM captioned at scale

The MONET dataset

Size	103.8M
License	Apache 2.0
Captions	Original alt-text + 4 VLM captions, short to long
Embeddings	DINOv2-ViT-g/14, CLIP ViT-B/32, SSCD, plus FAISS indexes
Annotations	YOLO-v9e detection, YOLO-v8x classification, MediaPipe faces
Latents	Pre-encoded with the SANA VAE

MONET curation pipeline

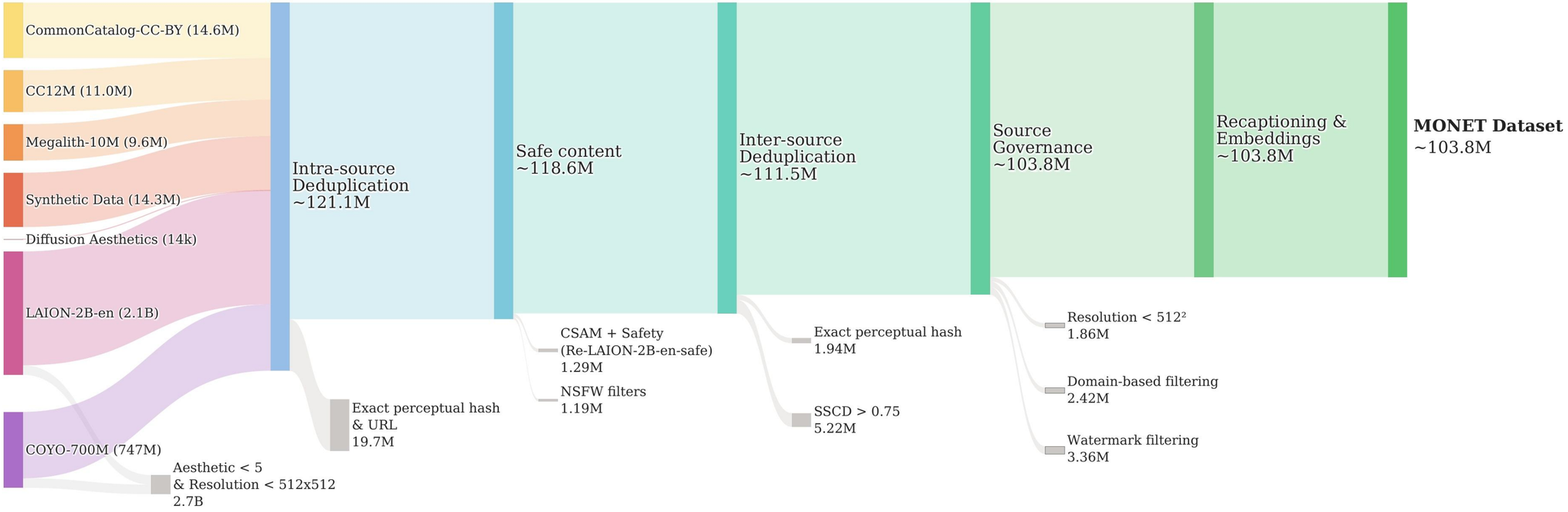
jasper

- 2.9B raw text-image pairs
- 6 sources with permissive licenses
- 14% synthetic data, generated with 3 open, top-performing, and Apache 2.0 T2I models
  - FLUX.1-schnell
  - FLUX.2-klein4B
  - Z-Image



(a) FLUX.1-schnell (b) FLUX.2-klein-4B (c) Z-Image

Examples of synthetic images in MONET.



Overview of the curation pipeline.

Pre-filtering & intra-source deduplication

jasper

2.9B images → 121.1M images

Pre-filtering: LAION and COYO

- Resolution > 512^2
- Aesthetic score > 5



3.76 4.09 4.71 5.34 6.26 7.51

Images with different aesthetic scores. A lower aesthetic score is correlated with visually unappealing images.

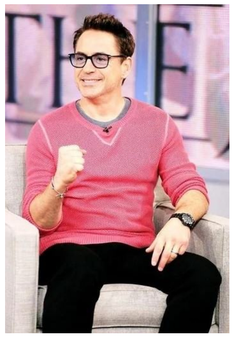
Intra-source deduplication: all sources

- **URL: exact copies**
- **pHash: near-exact copies** differing on compression and scaling
    1. Apply 2D Discrete Cosine Transform (DCT)
    2. Keep low frequencies
    3. Compare pairs of images with Hamming distance

Safety filtering

121.1M images → 118.6M images

- **Conservative approach**
- Restrict to Re-LAION Safe (−1.29M)
- **Ensemble of Not-safe-for-work (NSFW) detectors** (−1.19M)
  - Jasper Internal → score in [0, 1], threshold in 0.5
  - Bumble → score in [0, 1], threshold in 0.34
  - Falcon → binary score



Band 1 (retained).



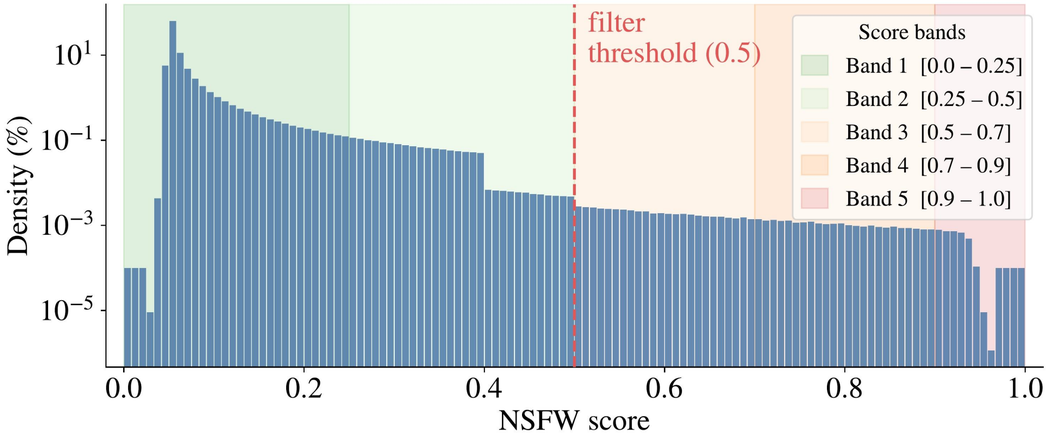
Band 2 (retained).



Band 3 (filtered).



Band 4 (filtered).



Jasper Internal NSFW detector.

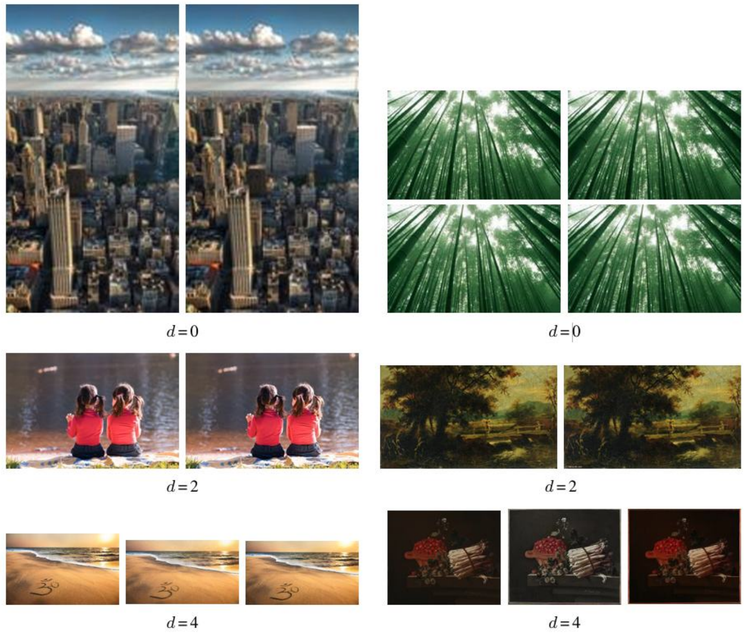
Inter-source deduplication

jasper

118.6M images → 111.5M images

1. URL and pHash (−1.94M)

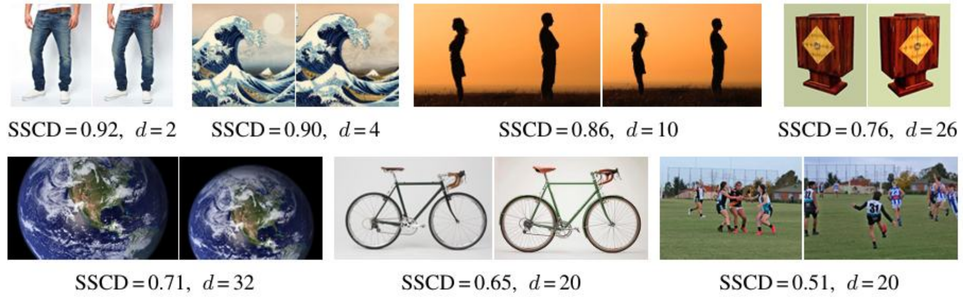
- Exact (URL) and near-exact (pHash) duplicates
- Catches copies that differ by compression or scaling



Examples of pHash duplicates.

2. Self Supervised Copy Detection (SSCD) (−5.22M)

- Near duplicates
- Compute 512-d SSCD embeddings
- Retrieve k=64 nearest neighbours (FAISS index)
- k=64: trades-off search speed and cluster recall, tuned empirically
- **Collapse pairs with cosine > 0.75 → keep the highest resolution and aesthetic score**



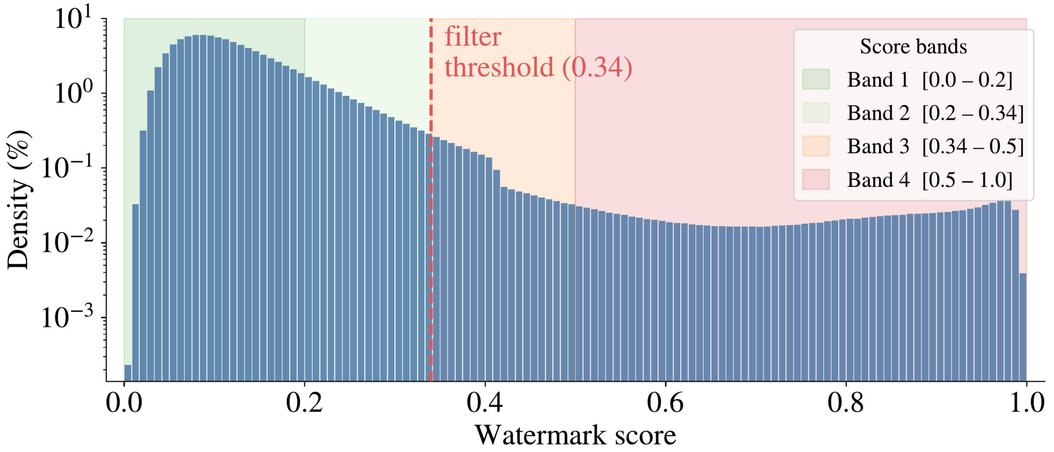
Examples of SSCD duplicates.

Domain filtering and governance

jasper

111.5M images → 103.8M images

1. Remove images with resolution < 512^2 (only LAION and COYO were pre-filtered)
2. Remove images from known list of stock photo providers (-2.42M)
   - Dreamstime (-1.5M)
   - Shutterstock (-225k)
   - Freepick (-176k)
   - Getty (-175k)
   - Unsplash (-170k)
   - etc.
3. Ensemble of watermark filters (-3.36M)
   - LAION
   - Jasper Internal

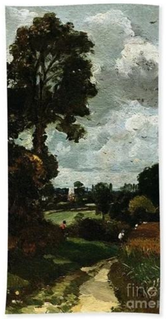




Band 1 (retained).



Band 2 (retained).



Band 3 (filtered).



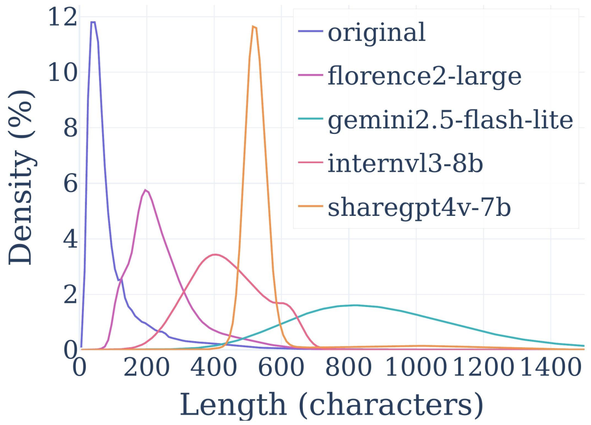
Band 4 (filtered).

Jasper Internal watermark detector.

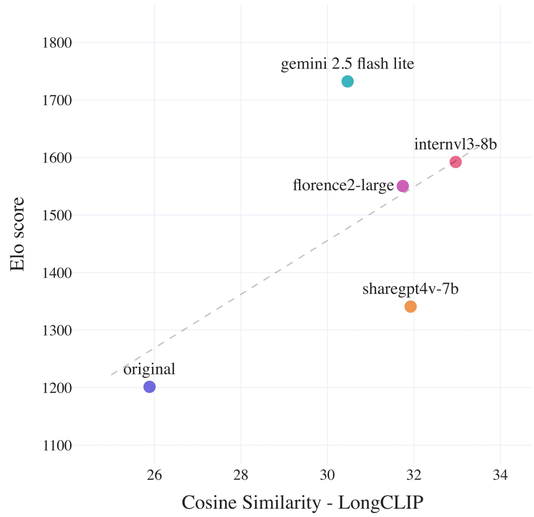
Multi-VLM re-captioning

jasper

- Caption quality and diversity are crucial for T2I
- **Rich and detailed captions** improve convergence
- Human-annotated captions are prohibitively expensive
- Fully re-caption MONET with:
  - Florence-2-Large → short, user-like captions
  - InternVL3-8B
  - ShareGPT4V-7B
  - Gemini 2.5 Flash-Lite → longest captions
- Longer and more detailed captions have higher Elo score and alignment with images (Long CLIP Score)



Caption length histogram.



Human voting (Elo score) and Long CLIP Score of different captions.

Dataset analysis

jasper

Image content

- Top-5 CLIP-based zero-shot classification
- Prompt: “an image of a {class}”
- 2.7k classes, grouped in 6 final classes
- **Classification done on all the dataset**
- **Consistent with comparable closed-source datasets (Qwen Image’s)**

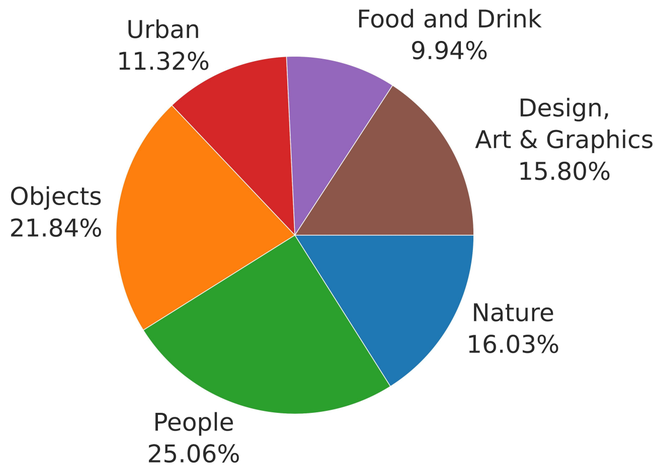


Image content distribution.

Image style

- Qwen3-VL-8B-Instruct classifies images on 15 classes
- **Classification done in a 1.5M random subset**

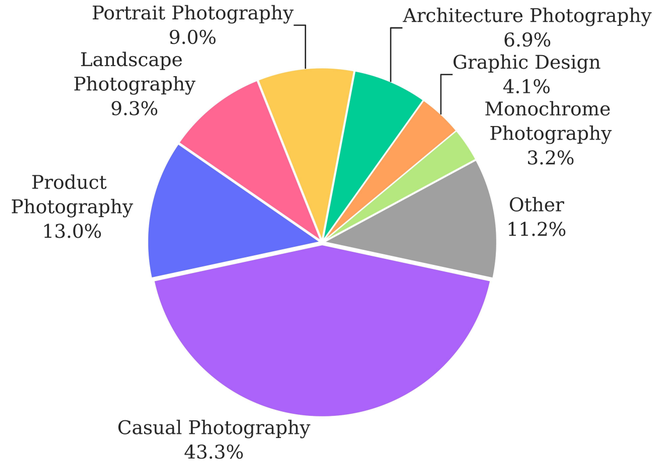


Image style distribution.

Downstream validation: T2I training

jasper

- Train a T2I Flow Matching model **exclusively on MONET**
- 4B MMDiT-based denoiser
- SANA Deep-compression VAE (DCVAE)
- Text conditioning: Qwen3-4B
- **Model is competitive with many models trained on closed data, and with additional fine-tuning and post-training**

Model	Num. Params (B)	GenEval							DPG					
		Objects		Counting	Colors	Position	Color	Overall ↑	Global	Entity	Attribute	Relation	Other	Overall↑
		Single	Two											
SD1.5 [78]	0.9	0.97	0.38	0.35	0.76	0.04	0.06	0.43	74.63	74.23	75.39	73.49	67.81	63.18
PixArt-α [11]	0.6	0.98	0.50	0.44	0.80	0.08	0.07	0.48	74.97	79.32	78.60	82.57	76.96	71.11
Emu3-Gen [100]	8.0	0.98	0.71	0.34	0.81	0.17	0.21	0.54	85.21	86.68	86.84	90.22	83.15	80.60
SDXL [70]	2.6	0.98	0.74	0.39	0.85	0.15	0.23	0.55	83.27	82.43	80.91	86.76	80.41	74.65
SD3 Medium [20]	2.0	0.98	0.74	0.63	0.67	0.34	0.36	0.62	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1 [Dev] [49]	12.0	0.98	0.81	0.74	0.79	0.22	0.45	0.66	74.35	90.00	88.96	90.87	88.33	83.84
DALL-E 3 [3]	–	0.96	0.87	0.47	0.83	0.43	0.45	0.67	90.97	89.61	88.39	90.58	89.83	83.50
SANA-1.5 [106]	4.8	0.99	0.85	0.77	0.87	0.34	0.54	0.72	–	–	–	–	–	85.00
Lumina-Image 2.0 [71]	2.6	–	0.87	0.67	–	–	0.62	0.73	–	**91.97**	90.20	**94.85**	–	87.20
Janus-Pro-7B [15]	7.0	0.99	0.89	0.59	0.90	**0.79**	0.66	0.80	86.90	88.90	89.40	89.32	89.48	84.19
HiDream-I1-Full [8]	17.0	**1.00**	**0.98**	0.79	0.91	0.60	0.72	0.83	76.44	90.22	89.48	93.74	91.83	85.89
Z-Image [110]	6.0	**1.00**	0.94	0.78	**0.93**	0.62	**0.77**	0.84	**93.39**	91.22	**93.16**	92.22	91.52	88.14
Qwen-Image [101]	20.0	0.99	0.92	**0.89**	0.88	0.76	**0.77**	**0.87**	91.32	91.56	92.02	94.31	**92.73**	**88.32**
Ours	4.1	**1.00**	0.90	0.73	0.88	0.35	0.62	0.74	84.80	91.76	89.70	94.16	79.60	85.56

Results on the GenEval and DPG benchmarks.



Example images generated by our T2I model trained exclusively on MONET.

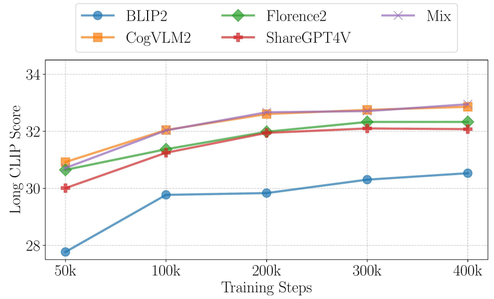
Downstream validation: ablations

jasper

Impact of multi-captioning

- Multiple captioners improve robustness and generalization
- Verbose captioners (ShareGPT4V) converge faster (lower FID) but hurt out-of-distribution performance when used alone

Captioner	FID ↓
BLIP2*	7.5
CogVLM2*	7.3
Florence2	7.3
ShareGPT4V	**5.2**
Mix	8.0



Captioner ablation (ImageNet), evaluated on validation set with all the captions.

Impact of synthetic data

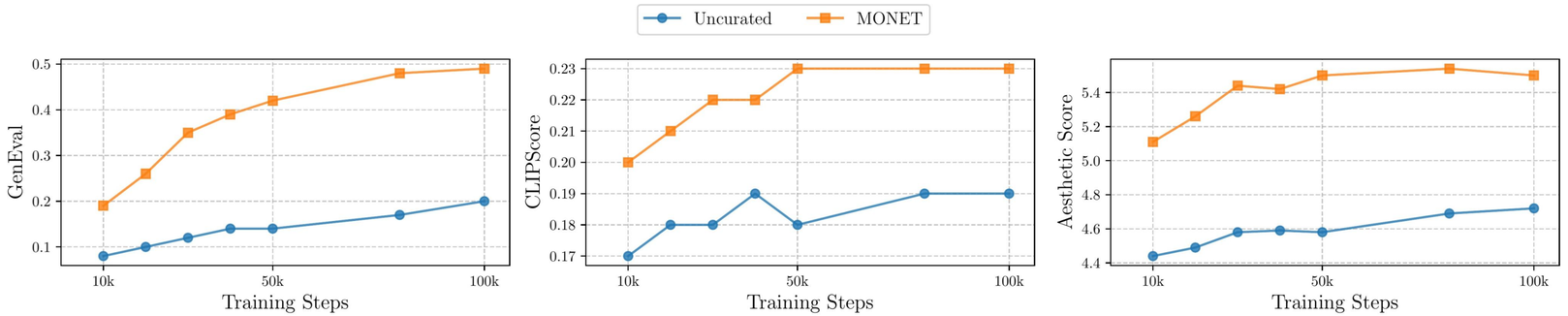
- Synthetic data improves image-text alignment (higher Long CLIP Score)
- 100% synthetic: higher FID, lower diversity (DIV-LPIPS)
- Adding 25% real data recovers diversity

Synthetic Data (%)	FID ↓	DIV-LPIPS ↑	LongCLIP ↑
0	8.1	0.59	32.95
10	8.3	0.59	33.32
25	8.0	0.59	33.44
50	7.6	**0.60**	**33.49**
75	**7.2**	0.59	33.44
100	15.0	0.47	34.21

Synthetic data ratio ablation (re-captioned ImageNet).

Impact of data curation

- Model trained on MONET outperforms model trained on 103.8M uncurated text-image pairs.



Training on MONET vs. uncurated 103.8M set.

Conclusions & limitations

jasper

Conclusions

- Apache 2.0 dataset of 103.8M curated image-text pairs
- First open, filtered, deduplicated and multi-captioned dataset for training T2I models at scale
- Validated by training a T2I exclusively on MONET, that reaches competitive GenEval and DPG scores
- Balanced image content & style distributions, consistent with comparable closed-source datasets (like Qwen Image's)

Limitations

- Inherits western biases from its sources
- English-only captions
- Image style and ethics audit
  - Done in a 1.5M representative subset
  - Rely on single VLM (Qwen3-VL-8B-Instruct)