Alchemist: Turning Public Text-to-Image Data into Generative Gold

Valerii Startsev[2,4], Alexander Ustyuzhanin[1], Alexey Kirillov[1, 3]
Dmitry Baranchuk[2], Sergey Kastryulin[2]

1 - Yandex
2 - Yandex Research
3 - Lomonosov Moscow State University
4 - Higher School of Economics

Motivation

SFT is actively used to improve the quality of T2I models. However, the success of SFT depends on the quality of the training dataset.

Existing approaches to creating datasets have drawbacks:

- **Proprietary Data:** Top-tier models are fine-tuned on closed-source, internal datasets, which slows progress for the open-source community.
- **Public Datasets Fall Short:** They are often too niche (e.g., anime) or filtered with simple heuristics that fail to identify truly impactful samples.
- **Human Curation is Ineffective:** Manually picking "good" images is expensive, slow, and surprisingly bad at predicting which samples will improve the model.

Question: *How to find samples that would maximally benefit post-SFT quality?*

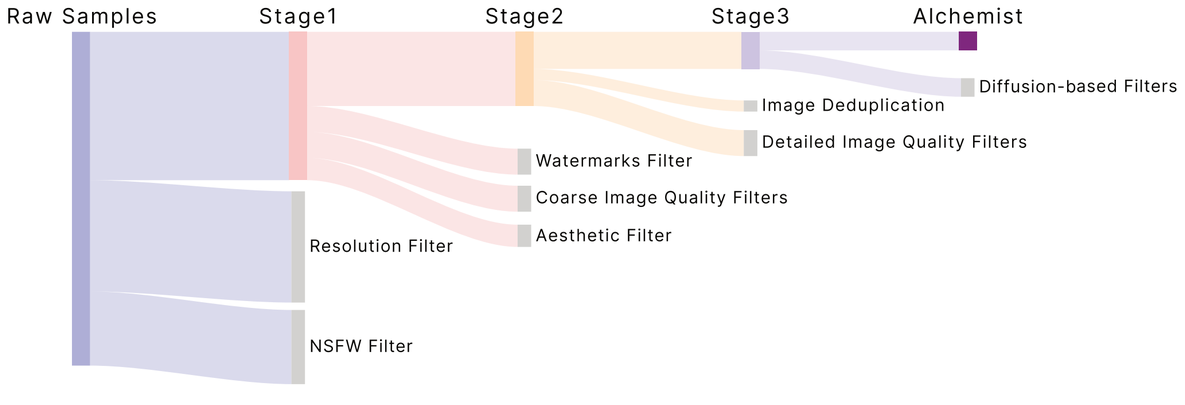
Answer: Leverage a pre-trained generative model as an estimator of high-impact training samples.

Contributions

- **Methodology:** A principled methodology for curating high-quality, general-purpose SFT datasets by leveraging a pre-trained generative model to identify samples that maximize post-SFT model improvement.
- **SFT Dataset:** Alchemist, a compact (3,350 samples) yet highly effective SFT dataset constructed via our methodology, significantly enhances text-to-image generation quality while maintaining output diversity and style.
- **Checkpoints:** Open-sourced, fine-tuned weights for five publicly available text-to-image models, demonstrating performance gains over their baselines after SFT with Alchemist.

Methodology

We use multi-stage image filtering pipeline, starting from *O(10 billion)* images, aggregated from web-scraped sources, and progressively filter it. Last stage leverages pre-trained diffusion model as an estimator for final filtering.



Last stage produces Alchemist dataset of 3,350 samples.

Huggingface Collection with paper, dataset and checkpoints

Diffusion-based Quality Estimator

Diffusion estimator feeds an image and a special prompt with quality-related keywords (like "aesthetic", "high quality") into a pre-trained diffusion model. It then calculates a quality score for the image by measuring the strength of the model's internal cross-attention activations corresponding to the most discriminative keywords and layers.

Algorithm 1: Diffusion-based Quality Estimator

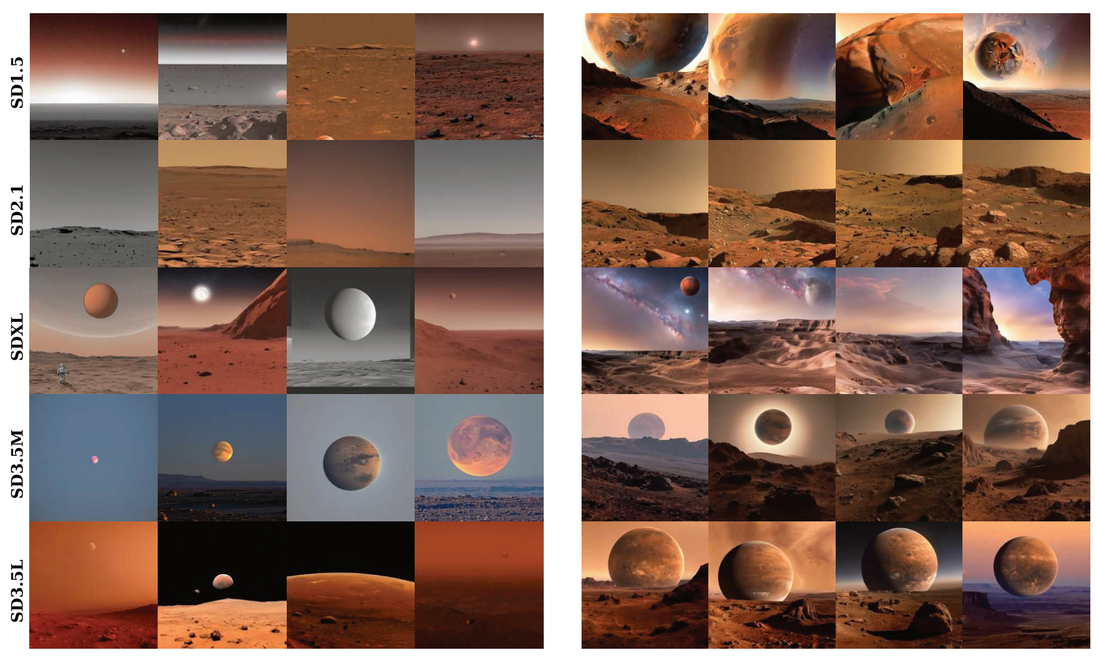
Input: X_{HQ}, X_{LQ}: Two groups of train images of higher and lower visual quality
X: Test images, $|X| = N$
ϵ_θ: Pretrained text-to-image generative model
$\mathcal{P}$: Predefined prompt with tokens $\{w_1, ..., w_M\}$
L: Number of cross-attention layers
K: Number of top discriminative features
t: Timestep for activation extraction
Output: Quality scores $\mathbf{f} \in \mathbb{R}^N$

1. **Extract activations:**
for *each image* $x \in \mathbb{R}^{h \times w}$ *in* $X_{HQ} \cup X_{LQ} \cup X$ **do**
  Save cross-attn maps $\{A^{(x)}_{l,m} \in \mathbb{R}^{h_l \times w_l}\}_{\substack{l=1...L \\ m=1...M}}$ during noise prediction via $\epsilon_\theta(x, \mathcal{P}, t)$
  Compute spatial activation norms:
    $N^{(x)}_{l,m} = \|A^{(x)}_{l,m,:,:}\|_2 \quad \forall l \in \{1, ..., L\}, m \in \{1, ..., M\}$
end
2. **Find (layer, token) pairs with most discriminative features:**
for *each* (l, m) *pair* **do**
  $s_{l,m} \leftarrow 0$
  for *each* $(x_{HQ} \in X_{HQ}, x_{LQ} \in X_{LQ})$ *pair* **do**
    Compute separation score:
      $s_{l,m} \mathrel{+}= \mathbb{I}[N^{(x_{HQ})}_{l,m} > N^{(x_{LQ})}_{l,m}]$
  end
end
Select top-K (l, m) pairs with highest $s_{l,m}$: $\mathcal{K} = \{(l_1, m_1), ..., (l_K, m_K)\}$
3. **Compute scores:**
for *each image* $x \in X$ **do**
  $\mathbf{f}_x = \sum_{(l,m) \in \mathcal{K}} N^{(x)}_{l,m}$
end
return *Quality scores* $\mathbf{f}$

Experiments

We fine-tuned five models from Stable Diffusion family on Alchemist dataset. For each model we found best hyperparameters using grid search over learning rate, number of training steps and other parameters.

Qualitative Results. Finetuning on Alchemist improves images in terms of aesthetic appeal, detail, and overall image complexity:



Quantitative Results. Human evaluation and automatic metrics confirm improvements of quality:

Model	Side-by-Side Win Rate				Automatic Metrics (Δ)			
	Rel.↑	Aes. ↑	Comp. ↑	Fidel. ↑	FD_{DINOv2} ↓	CLIP ↑	IR ↑	HPS-v2 ↑
SD1.5-Alchemist					129.8	0.277	**0.38**	**0.270**
vs baseline	0.53	0.64	0.78	0.47	131.5	0.279	0.02	0.243
vs LAION-tuned	0.47	0.60	0.73	0.45	**112.1**	**0.286**	0.32	0.260
SD2.1-Alchemist					**95.6**	0.281	0.62	**0.282**
vs baseline	0.57	0.69	0.81	0.56	129.3	0.276	0.18	0.253
vs LAION-tuned	0.49	0.56	0.72	0.52	112.4	**0.287**	**0.65**	0.278
SDXL-Alchemist					97.4	0.286	0.76	**0.292**
vs baseline	0.52	0.61	0.78	0.51	**73.4**	0.293	0.71	0.283
vs LAION-tuned	0.49	0.58	0.78	0.57	108.9	**0.294**	**0.81**	0.291
SD3.5M-Alchemist					**76.2**	0.286	**1.07**	**0.295**
vs baseline	0.51	0.57	0.67	0.50	81.4	**0.287**	0.97	0.292
vs LAION-tuned	0.48	0.58	0.73	0.49	87.9	0.286	0.87	0.274
SD3.5L-Alchemist					**80.9**	0.287	**1.12**	**0.299**
vs baseline	0.49	0.62	0.72	0.41	91.4	0.286	1.01	0.298
vs LAION-tuned	0.47	0.57	0.76	0.55	91.1	**0.297**	1.10	0.294

Rel - relevance, text-image correspondence; *Aes* - aesthetics;
Comp - complexity; *Fidel* - artifacts and distortions; *FD* - Fréchet Distance;
IR - ImageReward; *HPSv2* - human preference score reward