TASTE: TRAINING-AWARE DATA SELECTION FOR TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

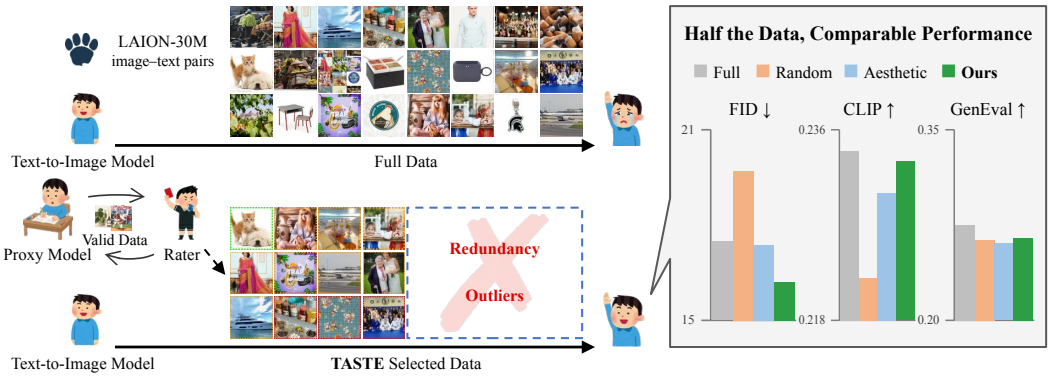


Figure 1: TASTE selects informative samples. Full uses all 30M data samples; Random uses a random 15M subset; Aesthetic uses the top 15M by Aesthetic score; and Ours uses 15M selected by TASTE. Valid data follows the same data distribution as the full data but is excluded from it.

ABSTRACT

The remarkable progress of text-to-image (T2I) generative models, such as Imagen, Stable Diffusion, and FLUX, has led to substantial improvements in visual quality. However, their performance is fundamentally limited by the quality of training data. Web-crawled and synthetic image datasets often contain low-quality or redundant samples which lead to degraded visual fidelity, unstable training, and inefficient computation. Hence, effective data selection is crucial for improving data efficiency. Existing approaches rely on costly manual curation or heuristic scoring based on single-dimensional features in text-to-image data filtering. Although meta-learning-based data selection has been explored for LLMs, its application to text-to-image model training remains underexplored. To this end, we propose **TASTE**, a meta-gradient-based framework to select a suitable subset from large-scale text-image data pairs. Our approach automatically estimates the influence of each sample by iteratively optimizing the model from a data-centric perspective. TASTE consists of two key stages: data rating and data pruning. We train a lightweight rater to estimate each sample's influence based on gradient information, enhanced with multi-granularity perception. We then use the Shift-Gsample strategy to select informative subsets. Experiments on both synthetic and web-crawled datasets demonstrate that TASTE consistently improves downstream model performance. Training on the 50% subset selected by TASTE achieves performance comparable to full-data training while improving data efficiency.

1 INTRODUCTION

The remarkable progress of text-to-image (T2I) generative models, such as Imagen (Saharia et al., 2022), Stable Diffusion (Esser et al., 2024), and FLUX (Labs, 2024), has marked a major milestone in artificial intelligence, enabling the synthesis of highly realistic and semantically aligned images from textual descriptions. These models have demonstrated transformative potential across diverse

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

domains, including art, design, and entertainment. However, the performance of such models is fundamentally bounded by the quality of their training data. While large-scale web-crawled (Schuhmann et al., 2022) and synthetic image datasets (Fang et al., 2025) provide abundant visual information, they inevitably contain low-quality samples—such as blurry images, plain-background advertisements, or other redundant content. Training on such noisy or less-informative data can lead to degraded visual fidelity, training instability, and inefficient resource consumption, significantly hindering the scalability of T2I model development (Abbas et al., 2023; Slyman et al., 2024). Effective data curation is therefore crucial for data efficiency (Albalak et al., 2024; Xie et al., 2023; Gu et al., 2024).

Existing data selection approaches remain insufficient. Manual data curation is prohibitively expensive and lacks scalability. Human-defined heuristic criteria (Penedo et al., 2024; Qiu et al., 2024; Wettig et al., 2024) often rely on classifiers that evaluate data from one single dimension. Although meta-learning-based methods (Zhuang et al., 2025; Calian et al., 2025) have been explored in Large Language Model training, recent data selection approaches for text-to-image model training still rely predominantly on surface-form features (LAION-AI, 2022), *e.g.*, CLIP-based (Radford et al., 2021) linear estimators that predict aesthetic or clarity scores. Such heuristic filters rely on predefined signals and do not directly optimize data selection for downstream T2I training performance.

To address these challenges, we propose **TASTE**, which is specially designed for large-scale text-image data selection via meta-gradient. Unlike heuristic scoring strategies, meta-learning simply requires users to specify *what* they want at the end of training. The model then autonomously learns *how* to value data samples through iterative meta-optimization, providing a scalable, data-driven approach to filter the training data stream. TASTE consists of two components: data rating and data pruning. We first train a lightweight, multi-granularity rater network to estimate each sample's influence from the training dynamics (*i.e.*, loss) observed during proxy-model training. We then propose the Shift-Gsample data pruning strategy to retain informative samples from the middle-to-late portion of the learned ranking. TASTE can also adapt across different data domains. We apply TASTE to both synthetic image datasets and web-crawled datasets and also use the selected datasets to train various model architectures. As shown in Fig. 1, TASTE achieves performance comparable to full-data training using only half of the data. Moreover, we find that the selected subsets also align well with human intuitions, effectively filtering out images with plain backgrounds or highly chaotic content.

2 Related Work

Coreset selection. Coreset selection (Phillips, 2016; Zheng et al., 2022) seeks a small, representative subset that yields performance comparable to full-data training. Many approaches rely on static heuristics. These include pruning data via predefined rules based on intrinsic properties (*e.g.*, resolution, watermarks) (Gururangan et al., 2020; Xia et al., 2020; Chen et al., 2023; Zheng et al., 2024; Gu et al., 2024), or using pretrained models to score proxy attributes like aesthetic quality or text-image alignment (Rombach et al., 2022; Yu et al., 2023; Wang et al., 2025). While effective for removing outliers, these strategies are limited because the proxy metrics are static and not directly optimized for the final downstream model's performance. To address this gap, recent methods rank data using gradients extracted once from a fixed warm-up model (Liu et al., 2024; Xia et al., 2024; Wu et al., 2024). MOTIVE (Wu et al., 2026) uses a shared timestep–noise pair $(t_{\text{fix}}, \epsilon_{\text{fix}})$ to stabilize influence estimates at lower cost. Such static scores cannot track changes in sample utility during training. In contrast, TASTE alternately updates the rater and proxy T2I model via bilevel meta-learning, adapting influence scores to the held-out validation objective. Alchemist (Startsev et al., 2025) and MOTIVE (Wu et al., 2026) both curate data for fine-tuning pretrained generative models.

Bilevel optimization for data selection. Model-based coreset selection methods (Xia et al., 2024; Zhuang et al., 2025; Calian et al., 2025) that compute the exact gradient of validation performance are often formulated as a bilevel optimization problem, which simultaneously optimizes both the rater and the inner model. While recent works (Calian et al., 2025; Nikdan et al., 2025; Dai et al., 2025) adopt second-order derivatives to find precise solutions, these methods are computationally expensive and thus infeasible for large-scale modalities such as text-to-image (T2I) models. To overcome this limitation, TASTE is instead closely related to penalty relaxation methods (Clarke, 1990; Shen & Chen, 2023; Liu et al., 2022; Kwon et al., 2023; Xiao et al., 2023; Lu, 2023; Shen

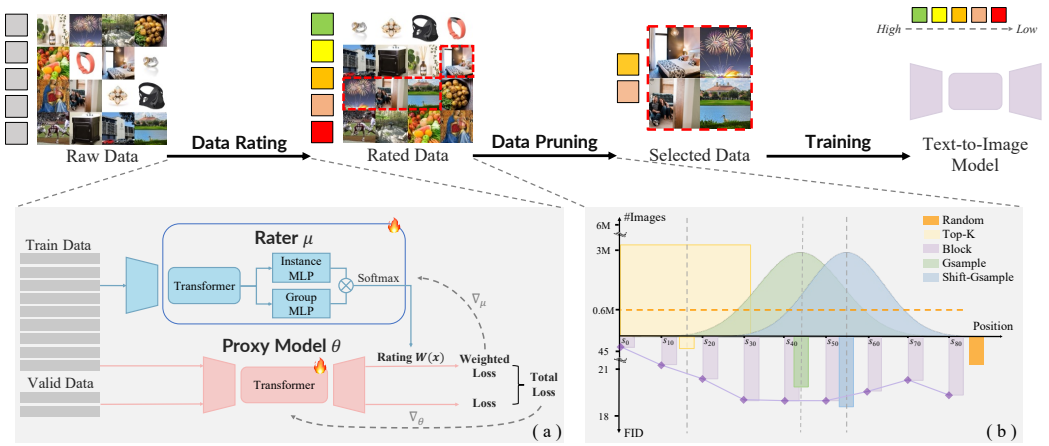


Figure 2: Overall pipeline of TASTE. In the initial data rating stage (a), the rater predicts an influence score for each image based on gradients extracted from a T2I proxy model. The rater and the proxy model are jointly optimized through weighted loss and total loss. In the data pruning stage (b), we introduce the Shift-Gsample strategy to efficiently retain informative samples while filtering out redundant data and outliers. The resulting TASTE-selected dataset enables highly efficient training of downstream text-to-image models.

et al., 2025), which efficiently approximate bilevel optimization solutions. Our work is most similar to (Shen et al., 2024), but we introduce a continuous scoring mechanism that evaluates each sample based on the alignment between its individual gradient and the validation gradient. This design facilitates highly efficient adaptation to new downstream tasks.

3 METHOD

Data selection for text-to-image training (Albalak et al., 2024; Xie et al., 2023; Gu et al., 2024) aims to select a small but influential subset that enables the model to achieve better or comparable performance. Instead of relying on hand-crafted heuristics, TASTE is optimized to directly maximize downstream model performance, ensuring that the chosen subset is well-suited for model learning. As illustrated in Fig. 2, TASTE proposes a validation-guided influence ranking for pruning and applies Shift-Gsample to construct an informative and diverse subset. The rating stage estimates sample influence through validation-guided meta-gradients, while the pruning stage converts the learned ranking into a subset under a specified data budget.

3.1 DATA RATING VIA META-LEARNING

A training sample is useful if learning from it also lowers the loss on a held-out validation set $\mathcal{D}_{\text{val}}$, which follows the training distribution but shares no samples with it (Calian et al., 2025; Pruthi et al., 2020). We therefore train a lightweight rater to predict, for each image–text pair, how much it helps a proxy T2I model reduce validation loss.

Formulation. Let θ denote a proxy T2I model and r_μ a rater with parameters μ. For each training sample $x_i \in \mathcal{D}_{\text{train}}$, the rater outputs a score $\hat{W}_{x_i} = r_\mu(x_i)$, which is normalized within a minibatch B into a training weight:

$$W_{x_i}(\mu) = \frac{\exp(\hat{W}_{x_i})}{\sum_{x_j \in B} \exp(\hat{W}_{x_j})}. \tag{1}$$

With per-sample loss $\ell_i(\theta) = \mathcal{L}(\theta; x_i)$, the proxy is trained on the weighted loss $F(\theta, \mu) = \mathbb{E}_B[\sum_{x_i \in B} W_{x_i}(\mu)\ell_i(\theta)]$, and evaluated by the validation loss $V(\theta) = \mathbb{E}_{x \sim \mathcal{D}_{\text{val}}}\mathcal{L}(\theta; x)$. Ideally, the rater should assign weights such that the resulting proxy generalizes best:

$$\mu^* = \arg\min_{\mu} V(\theta^*(\mu)), \qquad \theta^*(\mu) \in \arg\min_{\theta} F(\theta, \mu). \tag{2}$$

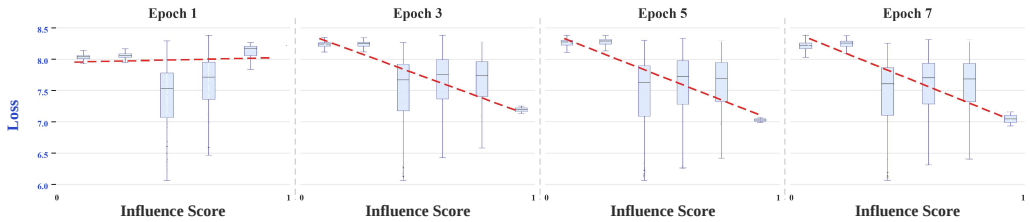


Figure 3: Loss across normalized influence-score ranges. For each training sample, we record its instantaneous loss during STAR-0.3B training and find that samples with lower loss generally receive higher influence scores once training has progressed sufficiently.

Solving this bilevel problem exactly requires retraining the proxy for every rater update, or differentiating through its whole training trajectory, which is infeasible at T2I scale.

Efficient rater optimization. We instead compare two copies of the proxy, both initialized from the same warmed-up checkpoint and fed the same minibatches. The *guided* proxy θ follows the weighted training loss plus the validation gradient, while the *reference* proxy $\hat{\theta}$ follows the weighted training loss only:

$$\begin{aligned}\theta_{k+1} &= \theta_k - \beta_k \left(g_{\text{train}}(\theta_k, \mu_k) + g_{\text{val}}(\theta_k)\right), \\ \hat{\theta}_{k+1} &= \hat{\theta}_k - \beta_k g_{\text{train}}(\hat{\theta}_k, \mu_k),\end{aligned} \tag{3}$$

where $g_{\text{train}} = \sum_{x_i \in B} W_{x_i} \nabla_\theta \ell_i$ and $g_{\text{val}} = \nabla_\theta V$. Since the two proxies differ only in the validation guidance, the per-sample loss gap $\Delta_{i,k} = \ell_i(\theta_k) - \ell_i(\hat{\theta}_k)$ measures how validation guidance changes the loss of x_i. A negative gap means that moving toward lower validation loss also makes x_i easier to fit, so x_i agrees with the validation objective. The rater is updated to increase the weights of such samples:

$$\mu_{k+1} = \mu_k - \alpha_k \sum_{x_i \in B} \Delta_{i,k} \nabla_\mu W_{x_i}(\mu_k). \tag{4}$$

All three updates use the current states and alternate throughout training (Algorithm 1); no unrolling or second-order derivatives are needed. This update has a principled origin: Eq. (4) follows from the value-gap penalty relaxation of Eq. (2) (Shen & Chen, 2023), $P(\theta, \mu) = V(\theta) + \lambda[F(\theta, \mu) - \min_{\theta'} F(\theta', \mu)]$ with $\lambda = 1$: approximating the inner minimizer by the reference proxy $\hat{\theta}$ gives $\partial_\mu P \approx \sum_{x_i \in B} \Delta_i \nabla_\mu W_{x_i}$. Locally, the gap also has a simple meaning. For one step of both proxies from a shared state u,

$$\ell_i(\theta^+) - \ell_i(\hat{\theta}^+) = -\beta \langle \nabla \ell_i(u), g_{\text{val}}(u) \rangle + O(\beta^2), \tag{5}$$

so a sample receives a higher weight when its gradient aligns with the validation gradient. Because of the softmax in Eq. (1), this comparison is relative: a sample gains weight when its gap is below the batch average. The learned score thus reflects how much a sample helps validation performance, rather than its intrinsic visual quality. As a relaxation with finite reference training, this procedure does not guarantee an exact bilevel solution.

Multi-granularity perception. Because minibatches can differ systematically in their overall quality, instance weights normalized within each batch are not directly comparable across batches. The Group MLP therefore calibrates this batch-level variation using the mean and variance of the pooled sample features. For pooled sample features $\{h_i\}_{i=1}^N$, the Instance MLP produces relative instance weights, while the Group MLP uses the feature mean and variance to produce a batch weight:

$$\begin{aligned}W_{\text{inst},i} &= \text{softmax}(\text{MLP}_{\text{inst}}(h_i)), \\ W_{\text{batch}} &= \sigma(\text{MLP}_{\text{group}}\left[\text{Mean}(\{h_i\}), \text{Var}(\{h_i\})\right]), \\ W_i &= W_{\text{inst},i} W_{\text{batch}}.\end{aligned} \tag{6}$$

The combined weight replaces W_{x_i} in the training objective and loss-gap update, so validation feedback trains both MLPs. This factorization captures relative instance influence and minibatch context.

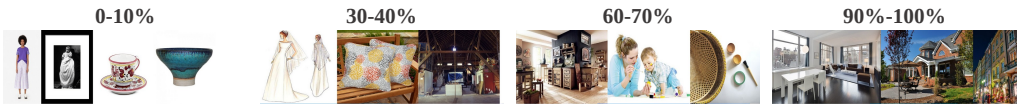


Figure 4: Image distribution across TASTE-ranked LAION subsets. Samples are ordered from high to low influence scores. The highest-ranked regions mainly contain visually simple images with plain backgrounds, the middle-to-late regions contain more content-rich and informative samples, and the tail increasingly contains noisy or cluttered scenes.

3.2 Data Pruning Strategy

Several data-selection methods for language model training first rate the samples and then apply Top-K pruning (Xia et al., 2024; Dai et al., 2025). For text, highly ranked samples are often clean and have low loss. These samples can still retain dense semantic information with limited token redundancy. For images, however, high-scoring samples may be visually plain. Such images are easy to fit and can rapidly reduce validation loss. Nevertheless, they may contain substantial spatial redundancy (Kersten, 1987). For example, large plain-background regions may provide little information relevant to the prompt. Directly transferring Top-K pruning to image data may therefore over-select visually plain samples.

Algorithm 1 Data rating with multi-granularity perception.

Require: $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$; proxies $\theta, \hat{\theta}$ from the same warm start; rater μ (Group + Instance MLP); step sizes α, β
Ensure: Trained rater μ^*
for $k = 0$ to K **do**
  Sample $B_k = \{x_i\}_{i=1}^{N} \subset \mathcal{D}_{\text{train}}$
  $h_i \leftarrow \text{Transformer}(x_i)$
  $\bar{h},\, v \leftarrow \text{Mean}(\{h_i\}),\, \text{Var}(\{h_i\})$
  $W_{\text{batch}} \leftarrow \sigma\big(\text{MLP}_{\text{group}}[\bar{h}, v]\big)$
  $W_{x_i} \leftarrow \text{softmax}\big(\text{MLP}_{\text{inst}}(h_i)\big) \cdot W_{\text{batch}}$
  $\theta_{k+1} \leftarrow \theta_k - \beta\big(g_{\text{train}}(\theta_k, \mu_k) + g_{\text{val}}(\theta_k)\big)$
  $\hat{\theta}_{k+1} \leftarrow \hat{\theta}_k - \beta\, g_{\text{train}}(\hat{\theta}_k, \mu_k)$
  $\Delta_{i,k} \leftarrow \ell_i(\theta_k) - \ell_i(\hat{\theta}_k)$
  $\mu_{k+1} \leftarrow \mu_k - \alpha \sum_{x_i \in B_k} \Delta_{i,k} \nabla_\mu W_{x_i}(\mu_k)$
end for
return $\mu^* = \mu_{K+1}$

Fig. 4 visualizes images from different influence-score ranges. The learned ranking contains a visually plain head, an informative middle-to-late region, and an increasingly chaotic tail. To further analyze these regions, we conduct a pilot study on 10k randomly sampled image–text pairs across the ranking. During proxy-model training, we track their per-sample loss $\mathcal{L}_i(t)$. Fig. 3 shows distinct loss distributions across the influence-score regions. In particular, the high-score region contains samples that are easier to fit. The downstream results in Tab. 4 further show that the middle-to-late region yields the strongest performance. Based on these observations, we introduce Shift-Gaussian sampling (**Shift-Gsample**; Fig. 2(b)) to focus data selection on this region.

It first discards the top $n\%$ region from the influence-ranked dataset. For each remaining sample x_i, let $r_i \in [n, 100]$ denote its original rank percentile and $u_i = 100(r_i - n)/(100 - n) \in [0, 100]$ its re-normalized percentile within the remaining sampling pool. We then sample the remaining data according to

$$p(i) \propto \exp\left(-\frac{(u_i - c)^2}{2\sigma^2}\right), \tag{7}$$

where c specifies the preferred percentile within the remaining sampling pool and σ controls the sampling spread. Given a target retention ratio ρ, we sample $\rho|\mathcal{D}_{\text{train}}|$ distinct examples without replacement from the remaining pool, which requires $\rho \leq 1 - n/100$. This concentrates the selected subset on the informative middle-to-late region while preserving diversity.

4 Experiments

4.1 Experimental Setup

Datasets. We use LAION-30M (Schuhmann et al., 2022), a broad but noisy web-crawled corpus and our default source; Flux-reason-6M (Fang et al., 2025), a FLUX-generated dataset with reasoning-oriented captions; and HPDv3 (Ma et al., 2025), which combines real and synthetic images in 1.17M

Table 1: Data selection on MJHQ-30K (Li et al., 2024) and GenEval (Ghosh et al., 2023) after three epochs from scratch. Ours-small uses fewer TASTE-selected images.

Method	#Img.	Time (h)	MJHQ-30K FID↓	MJHQ-30K CLIP↑	GenEval Score↑
Full	30M	65.34	17.48	0.234	**0.2752**
Random	15M	34.60	19.70	0.222	0.2632
Aesthetic	15M	34.60	<u>17.36</u>	<u>0.230</u>	0.2604
Clarity	15M	34.60	17.85	0.226	0.2251
Frequency	15M	34.60	18.77	0.228	0.2519
Edge-density	15M	34.60	20.13	0.224	0.2429
Ours-small	6M	13.08	18.22	0.228	0.2367
Ours	15M	34.60	**16.20**	**0.233**	<u>0.2645</u>

preference pairs, including outputs from older models. Valid data is a disjoint random split from the same source, comprising less than 5% of its samples.

Evaluation metrics. Following the evaluation protocol of our base model, STAR-T2I (Ma et al., 2024), we evaluate the trained models on MJHQ-30K (Li et al., 2024) using FID for image fidelity and CLIP-Score for image–text alignment, and use GenEval (Ghosh et al., 2023) to assess generation capabilities in complex, multi-object scenarios. See the supplementary material for evaluation details.

Models for data selection and training. We train scale-wise autoregressive STAR-40M, STAR-0.3B, and STAR-0.9B (Ma et al., 2024) from scratch and fine-tune FLUX-mini-3B (ARC, 2024) with LoRA at their corresponding optimal resolutions. FLUX-mini is distilled from FLUX-dev (Labs, 2024) and pruned in depth. The rater uses the STAR-40M backbone for features; the proxy uses a pretrained STAR-0.3B checkpoint. Downstream STAR models in Tab. 2(a) share the architecture but not the proxy weights.

Training budget. With 1M valid samples, 2–3 epochs of rater training take approximately 2.5 hours. We train a separate rater per source, score that source once, and prune the scores in a few minutes. Reusing the ranking and subsets across downstream architectures and scales amortizes this one-time cost (Tab. 2(a)).

Implementation details. All experiments are implemented in PyTorch on NVIDIA A100 GPUs. The rater model is trained with a batch size of 64 across 2 GPUs, while the downstream text-to-image models are trained with the maximum feasible batch size across 8 GPUs. For fairness, we keep the batch size consistent across all comparison methods with the same model. For evaluating different data-selection strategies, we use STAR-0.3B. STAR models are trained from scratch for three epochs, and FLUX-mini is fine-tuned for one epoch unless otherwise stated. See the supplementary material for compute, hyperparameter, and algorithm details.

4.2 Main Results

We present comparisons with other approaches in Tab. 1, main results of TASTE across different models in Tab. 2(a) and adaptation to different datasets in Tab. 2(b).

Comparable performance with half the data. We compare our TASTE with several methods, including using the full dataset (Full), random sampling (Random), and heuristic selection metrics based on four common image quality metrics—Aesthetic and Clarity (LAION-AI, 2022), Frequency (Castleman, 1979), and Edge Density (Otsu et al., 1975). For a fair comparison, the pruning strategy for each heuristic ranking is selected using the validation study in the supplementary.

As shown in Tab. 1, the selected 50% subset improves FID from 17.48 to 16.20, while CLIP (0.233 vs. 0.234) and GenEval (0.2645 vs. 0.2752) remain close to the full-data baseline. This supports comparable overall performance with half the training data, rather than uniform improvement on every metric. Moreover, using merely 20% of the data selected by TASTE yields performance comparable to training with 50% randomly sampled data, demonstrating substantially improved data efficiency. The qualitative comparison in Fig. 6 further shows that, under the same 15M-data

Table 2: Generalization across models, data scales, domains, and training budgets. (a) Green superscripts give improvements over random selection; STAR is trained from scratch and FLUX-mini uses LoRA. (b) Retention ratios on HPDv3-2M and Flux-reason-6M. (c) Matched iterations: Full uses 5 epochs; 50% subsets use 10 epochs.

(a) Transfer across model and data scales

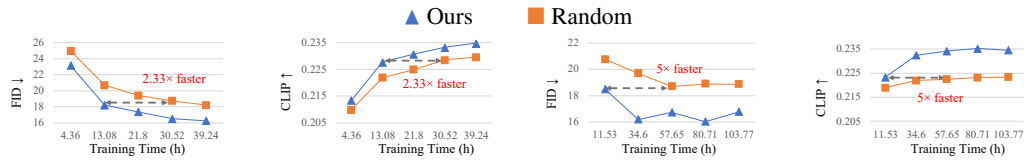
Setting	Model	#Params	#Images	FID↓	CLIP↑
Model scale	STAR	40M	6M	$26.74^{-1.92}$	$0.206^{+0.005}$
	STAR	0.3B	6M	$18.22^{-2.47}$	$0.228^{+0.006}$
	STAR	0.9B	6M	$16.71^{-1.43}$	$0.236^{+0.003}$
Data scale	STAR	0.3B	3M	$21.54^{-0.61}$	$0.219^{+0.002}$
	STAR	0.3B	6M	$18.22^{-2.47}$	$0.228^{+0.006}$
	STAR	0.3B	15M	$16.20^{-3.50}$	$0.233^{+0.011}$
Model family	FLUX-mini	3B	6M	$19.98^{-0.43}$	$0.237^{+0.001}$

(b) Transfer across data domains

Data	HPDv3		Flux-reason	
	FID ↓	CLIP ↑	FID ↓	CLIP ↑
Random 20%	35.55	0.198	23.66	0.214
Aesthetic 20%	33.78	0.200	25.38	0.207
Ours 20%	**32.27**	**0.204**	**22.78**	**0.214**
Random 50%	20.21	0.220	19.35	0.226
Aesthetic 50%	19.98	0.218	19.27	0.223
Ours 50%	**18.15**	**0.227**	**18.59**	**0.227**

(c) Matched training iterations

Dataset	Method	FID↓	CLIP↑
LAION	Full train (30M)	16.50	0.234
	Random (15M)	17.09	0.223
	Ours (15M)	**16.70**	**0.235**
Flux-Reason	Full train (6M)	15.80	0.237
	Random (3M)	15.82	0.235
	Ours (3M)	**14.38**	**0.236**



(a) FID, 6M (b) CLIP, 6M (c) FID, 15M (d) CLIP, 15M

Figure 5: Training efficiency on MJHQ-30K under matched data budgets. TASTE reaches random-selection performance 2.33× faster with 6M samples and 5× faster with 15M samples.

budget, TASTE produces more faithful object structures and richer visual details than Aesthetic-score selection, while achieving visual quality comparable to the Full model trained with twice as many samples (30M).

Effectiveness across models and scales. Tab. 2(a) evaluates transfer of selected subsets to autoregressive STAR and flow-based FLUX-mini. TASTE consistently outperforms random selection across model and data scales: a ranking learned with the small proxy improves STAR-0.3B and STAR-0.9B from-scratch training as well as FLUX-mini-3B LoRA fine-tuning. Tab. 6 further covers other retention ratios.

Scalability and reuse. Rater training is a one-time cost for each source dataset. The resulting ranking and subsets are reused across downstream model scales and families, amortizing this cost.

Adaptability across different data domains. TASTE also generalizes effectively across diverse data sources. We conduct experiments on HPDv3 (Ma et al., 2025) and Flux-reason (Fang et al., 2025), comprising 2M and 6M image-text pairs respectively. HPDv3 mixes real and synthetic images with human-preferred and non-preferred pairs, while Flux-reason is fully synthetic and reasoning-oriented. Tab. 2(b) evaluates 20% and 50% as representative retention ratios. At both ratios, TASTE consistently achieves better overall performance than random sampling and Aesthetic-score selection on HPDv3-2M and Flux-reason-6M. Together with the web-crawled LAION results, this shows that TASTE adapts well to heterogeneous data domains.

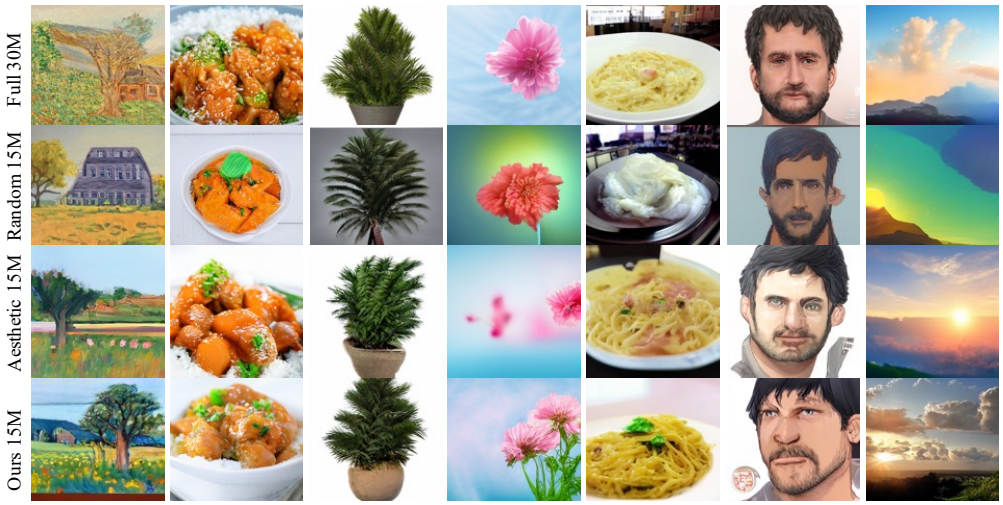


Figure 6: Qualitative comparison of models trained from scratch using different LAION training sets. Full uses all 30M samples; Aesthetic uses 15M samples selected by aesthetic score; Random uses 15M randomly sampled examples; and TASTE uses a 15M subset selected by our framework.

Table 3: Consolidated ablations of (a) pruning on 6M images, (b) validation-set size, and (c) Shift-Gsample sampling center c and spread σ. Cells in (c) report FID / CLIP. Bold and underlined values mark the best and second-best results within each study.

(a) Pruning on 6M images

Strategy	FID↓	CLIP↑
TopK	48.20	0.195
Gsample	19.22	0.225
Shift-Gsample	<u>18.37</u>	<u>0.227</u>
+ Group-MLP	**18.22**	**0.228**

(b) Validation-set size

Valid size	Time (h)	FID↓	CLIP↑
0.1M	∼0.33	<u>16.47</u>	0.230
0.5M	∼1.17	16.76	**0.234**
1M	∼2.50	**16.20**	<u>0.233</u>

Rater training time on two A100 GPUs.

(c) Shift-Gsample sampling center and spread

Center c	$\sigma = 6$	$\sigma = 8$	$\sigma = 10$
66.7%	18.54 / 0.229	19.16 / 0.228	18.99 / 0.226
50.0%	**16.20 / 0.233**	<u>16.71 / 0.231</u>	17.23 / 0.230

Comparison under equal training iterations. With equal iterations, Tab. 2(c) shows that TASTE beats random sampling and stays comparable to full-data training on LAION and Flux-Reason.

Efficiency in text-to-image model training. As shown in Fig. 5, TASTE-selected data enable the text-to-image model STAR to match the performance of the corresponding random-subset models 5× earlier in downstream training at 50% retention and 2.33× earlier at 20% retention.

Training trends. Fig. 5 tracks models trained on 6M and 15M samples. TASTE improves performance faster and more steadily than random sampling, which shows early fluctuations and slower gains; both fluctuate near convergence.

4.3 Ablation Study

Valid data size. Tab. 3(b) reports rater training time and downstream performance for three valid-data sizes. Their comparable results show that the rater does not require extensive validation data.

Multi-granularity perception and pruning strategies. Due to memory constraints, we perform gradient updates within minibatches rather than on the entire dataset. Differences between batches can introduce bias. To mitigate this, we incorporate a group MLP in the rater to capture global

Table 4: Performance of pruning strategies across different influence-score regions. Gray rows indicate the better-performing regions.

Method	#Img.	Range	FID↓	CLIP↑
BLOCK (s0)	6M	0%–20%	48.20	0.194
BLOCK (s10)	6M	10%–30%	21.21	0.221
BLOCK (s20)	6M	20%–40%	20.90	0.221
BLOCK (s30)	6M	30%–50%	19.80	0.222
BLOCK (s40)	6M	40%–60%	19.77	0.222
BLOCK (s50)	6M	50%–70%	19.70	0.222
BLOCK (s60)	6M	60%–80%	19.83	0.223
BLOCK (s70)	6M	70%–90%	21.30	0.220
BLOCK (s80)	6M	80%–100%	20.42	0.221
Random	15M	–	19.70	0.222
Full	30M	0%–100%	<u>17.48</u>	**0.234**
Ours	6M	–	18.22	0.228
Ours	15M	–	**16.20**	<u>0.233</u>

information across the dataset. We compare the pruning strategies and the contribution of Group-MLP in Tab. 3(a).

All strategies in Tab. 3(a) use the learned ranking. Shift-Gsample outperforms Top-K and Gsample by selecting informative middle-to-late regions; Group-MLP further improves results through minibatch context. The gain also depends on the ranking: supplementary experiments show that Shift-Gsample underperforms Top-K on Aesthetic and Clarity rankings. Figs. 3 and 4 characterize the regions' loss dynamics and images; supplementary results track score evolution during training. The pruning strategy also adapts across data domains.

Shift-Gsample hyperparameters. We ablate the sampling center c and standard deviation σ used by Shift-Gsample. Tab. 4 favors middle-to-late ranges over the highest-ranked region, motivating the fixed cutoff $n = 25$. We therefore test sampling centers $c \in \{50, 66.7\}$ and standard deviations $\sigma \in \{6, 8, 10\}$.

Results in Tab. 3(c) show that $c = 50$ and $\sigma = 6$ perform best, while shifting the sampling center toward the tail degrades performance. It also gives the best GenEval, FID, and CLIP among the 15M subsets in Tab. 1, while staying competitive with full-data training.

5 Analysis

What are the differences in image modality? Unlike text, images can contain extensive spatial redundancy, so visually clean, low-loss samples are not necessarily the most informative for T2I training. TASTE therefore proposes an influence ranking to guide pruning, favoring informative samples while retaining limited easy and noisy data for diversity.

Why autoregressive proxy T2I model? Although our proxy model is autoregressive, the resulting ranking transfers to the evaluated autoregressive STAR models and the flow-based FLUX-mini model, as shown in Tab. 2(a). We adopt an AR proxy primarily for its single-step generation process, which enables efficient gradient propagation and stable meta-optimization. Diffusion objectives typically sample both a denoising timestep and a noise realization, causing per-sample gradients to vary with these stochastic choices. The autoregressive proxy avoids this extra sampling and provides a more stable signal for meta-gradient estimation.

6 Conclusion

TASTE introduces meta-gradient data selection to T2I training. Under the same data budget, TASTE-selected subsets consistently outperform random sampling, while the 50% subset remains comparable to full-data training. **Limitation.** Due to limited compute, validating TASTE-selected subsets on much larger T2I models trained from scratch is left to future work.

AI USE STATEMENT

We used AI tools to polish the language and improve the readability of this manuscript. All AI-assisted revisions were reviewed and verified by the authors, who take full responsibility for the final content and its compliance with academic integrity standards.

REPRODUCIBILITY STATEMENT

We provide the source code in the supplementary material. The method is described in Section 3, and the experimental setup is given in Section 4.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

Tencent ARC. Fluxkits: A repo that facilitates the usage of flux series models, 2024. URL `https://github.com/TencentARC/FluxKits`. Accessed: 2025-11-09.

Dan A Calian, Gregory Farquhar, Iurii Kemaev, Luisa M Zintgraf, Matteo Hessel, Jeremy Shar, Junhyuk Oh, András György, Tom Schaul, Jeffrey Dean, et al. Datarater: Meta-learned dataset curation. *arXiv preprint arXiv:2505.17895*, 2025.

Kenneth R Castleman. *Digital image processing*. Prentice Hall Professional Technical Reference, 1979.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36:36000–36040, 2023.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Yalun Dai, Yangyu Huang, Xin Zhang, Wenshan Wu, Chong Li, Wenhui Lu, Shijie Cao, Li Dong, and Scarlett Li. Data efficacy for language model training. *arXiv preprint arXiv:2506.21545*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. Data selection via optimal control for language models. *arXiv preprint arXiv:2410.07064*, 2024.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Daniel Kersten. Predictability and redundancy of natural images. *Journal of the Optical Society of America A*, 4(12):2395–2400, 1987.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. *arXiv preprint arXiv:2309.01753*, 2023.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

LAION-AI. aesthetic-predictor: A linear estimator on top of clip to predict the aesthetic quality of pictures. GitHub repository, 2022. URL `https://github.com/LAION-AI/aesthetic-predictor`.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems*, 35: 17248–17262, 2022.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. Less is more: High-value data selection for visual instruction tuning. *arXiv preprint arXiv:2403.09559*, 2024.

Songtao Lu. Slm: A smoothed first-order lagrangian method for structured constrained nonconvex optimization. *Advances in Neural Information Processing Systems*, 36:80414–80454, 2023.

Xiaoxiao Ma, Mohan Zhou, Tao Liang, Yalong Bai, Tiejun Zhao, Biye Li, Huaian Chen, and Yi Jin. Star: Scale-wise text-conditioned autoregressive image generation. *arXiv preprint arXiv:2406.10797*, 2024.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Mahdi Nikdan, Vincent Cohen-Addad, Dan Alistarh, and Vahab Mirrokni. Efficient data selection at scale via influence distillation. *arXiv preprint arXiv:2505.19051*, 2025.

Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11 (285-296):23–27, 1975.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Jeff M Phillips. Coresets and sketches. *arXiv preprint arXiv:1601.00617*, 2016.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Jiantao Qiu, Haijun Lv, Zhenjiang Jin, Rui Wang, Wenchang Ning, Jia Yu, ChaoBin Zhang, Zhenxiang Li, Pei Chu, Yuan Qu, et al. Wanjuan-cc: A safe and high-quality open-sourced english webtext dataset. *arXiv preprint arXiv:2402.19282*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International conference on machine learning*, pp. 30992–31015. PMLR, 2023.

Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection. *arXiv preprint arXiv:2410.07471*, 2024.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and rlhf. *Journal of Machine Learning Research*, 26(114):1–49, 2025.

Eric Slyman, Stefan Lee, Scott Cohen, and Kushal Kafle. Fairdedup: Detecting and mitigating vision-language fairness disparities in semantic dataset deduplication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13905–13916, 2024.

Valerii Startsev, Alexander Ustyuzhanin, Alexey Kirillov, Dmitry Baranchuk, and Sergey Kastryulin. Alchemist: Turning public text-to-image data into generative gold. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.19297.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8428–8437, 2025.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*, 2024.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*, 2024.

Xindi Wu, Despoina Paschalidou, Jun Gao, Antonio Torralba, Laura Leal-Taixé, Olga Russakovsky, Sanja Fidler, and Jonathan Lorraine. Motion attribution for video generation. *arXiv preprint arXiv:2601.08828*, 2026.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. Predicting performance for natural language processing tasks. *arXiv preprint arXiv:2005.00870*, 2020.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.

Quan Xiao, Songtao Lu, and Tianyi Chen. A generalized alternating method for bilevel learning under the polyak-{\L} ojasiewicz condition. *arXiv preprint arXiv:2306.02422*, 2023.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36: 34201–34227, 2023.

Jia Yu, Lichao Zhang, Zijie Chen, Fayu Pan, MiaoMiao Wen, Yuming Yan, Fangsheng Weng, Shuai Zhang, Lili Pan, and Zhenzhong Lan. Quality and quantity: Unveiling a million high-quality images for text-to-image synthesis in fashion design. *arXiv preprint arXiv:2311.12067*, 2023.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. *arXiv preprint arXiv:2210.15809*, 2022.

Haizhong Zheng, Elisa Tsai, Yifu Lu, Jiachen Sun, Brian R Bartoldson, Bhavya Kailkhura, and Atul Prakash. Elfs: Label-free coreset selection with proxy training dynamics. *arXiv preprint arXiv:2406.04273*, 2024.

Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. Meta-rater: A multi-dimensional data selection method for pre-training language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10856–10896, 2025.

Outline. This supplementary material provides additional implementation details, experimental analyses, qualitative results, prompts, and limitations supporting the main manuscript.

1. **Section A: Additional Experimental Setup**
   - Section A.1: Evaluation Metrics
   - Section A.2: Compute Resources
   - Section A.3: Training Datasets
   - Section A.4: Baseline Details
2. **Section B: Training Details**
   - Section B.1: Hyperparameters
   - Section B.2: Rating Procedure and Score Dynamics
3. **Section C: Detailed Experimental Results**
   - Section C.1: Quantitative Results
   - Section C.2: Qualitative Results
   - Section C.3: Prompt Lists
4. **Section D: Limitations and Future Work**

A Additional Experimental Setup

A.1 Evaluation Metrics

For general evaluation, we assess our trained models on a suite of standard image-generation benchmarks, including MJHQ-30K (Li et al., 2024) and GenEval (Ghosh et al., 2023). MJHQ-30K automatically evaluates both aesthetic quality and semantic alignment. We report the **Fréchet Inception Distance (FID)** (Heusel et al., 2017) to quantify the distributional discrepancy between generated and real images, which serves as a proxy for perceptual fidelity. In addition, we compute the **CLIP Score** (Radford et al., 2021) to measure text–image semantic consistency, using the official CLIP ViT-B/32 model for both image and text encoding. GenEval is an object-centric evaluation framework focusing on compositional properties such as object co-occurrence, spatial arrangement, count accuracy, and color correctness. Since our from-scratch models are trained using lower-quality data at a smaller resolution (256×256) and with fewer parameters, their absolute performance is naturally lower than that of the larger base models.

A.2 Compute Resources

We train the rater using 2 NVIDIA A100 (80GB) GPUs for 2–3 epochs. With 1M valid samples, rater training takes approximately 2.5 hours. After training, we score each source dataset in a single pass using 8 A100 GPUs. Rater training and full-dataset scoring together take less than 3 hours of wall-clock time for LAION-30M, approximately 1.5 hours for Flux-reason-6M, and approximately 1 hour for HPDv3-2M. The subsequent pruning step scans the resulting scores once and takes only a few minutes. These one-time data-selection costs exclude downstream-model training. All text-to-image models (40M, 0.3B, 0.9B, and 3B parameters) trained on datasets of size 2M, 6M, and 30M use 8 A100 GPUs.

A.3 Training Datasets

Our experiments cover a diverse set of text–image pairs, broadly categorized into web-crawled and synthetic sources (see Fig. 7). Web-crawled datasets offer richer content and higher fidelity but are noisy and have challenging caption matching. Synthetic datasets have higher quality, clearer subjects, and better alignment, but may appear artificial with less realistic details. **LAION**-30M (Schuhmann et al., 2022). We randomly sample 30M images from LAION-100M, with captions recaptioned using CogVLM. **Flux-reason**-6M (Fang et al., 2025). A 6M synthetic dataset generated by the state-of-the-art FLUX-1.dev model, paired with GCoT-style captions that not only describe the scene but also articulate underlying reasoning and compositional intent. **HPDv3**-2M (Ma et al.,

2025). HPDv3 contains 1.17M preference pairs combining high-quality real images, synthetic images from recent generative models, and data inherited from HPDv2 and JourneyDB. We expand the preferred and unpreferred images in each pair into individual image–text samples, resulting in approximately 2M training samples.

Valid-data protocol. For each source dataset, we randomly hold out approximately 3.3% of its samples as valid data. The valid split follows the same source distribution as the candidate training pool but remains disjoint from it. We train a separate rater for each source dataset to learn its distribution-specific characteristics. The resulting ranking is not transferred between source datasets; instead, each selected subset can be reused across the downstream model scales and architectures evaluated for that source dataset.

Distribution across rating regions. Figure 4 in the main paper visualizes samples from different percentiles of the TASTE-ranked LAION dataset. The top-ranked subset (0–20%) mainly contains clean and simple images, often with plain backgrounds; although easy to fit, these samples provide limited semantic coverage. The middle-to-late region (40–80%) contains more informative samples with clear subjects and well-defined actions. The lowest-ranked region (90–100%) is dominated by noisy, cluttered, or semantically inconsistent samples. This distribution motivates Shift-Gsample, parameterized by sampling center c and spread σ, to favor the informative middle-to-late region rather than directly retaining only the highest-scoring samples.

A.4 Baseline Details

We evaluate our method against a diverse set of baseline sampling strategies. Full Training uses the entire dataset without any filtering. Random Sampling selects data uniformly under a fixed budget. The Aesthetic ranking baseline orders data by predicted aesthetic quality (LAION-AI, 2022), while the Clarity ranking (LAION-AI, 2022) baseline sorts samples by perceptual sharpness. The Edge-density ranking strategy measures structural complexity by computing grayscale edge maps and using the proportion of edge pixels as the sorting criterion. The Frequency ranking baseline orders samples by computing the ratio of low-frequency energy using FFT, assigning higher priority to images with a larger proportion of low-frequency components.

Baseline scope. To our knowledge, existing learned or meta-gradient data-selection methods are primarily designed for language-model training and do not provide a directly applicable pipeline for selecting image–text data for T2I models trained from scratch. Adapting them would require modality-specific proxy objectives and scoring architectures, which would confound a direct comparison. We therefore compare against random sampling and representative image-based rankings based on aesthetics, clarity, frequency, and edge density under the same candidate pool, data budget, downstream model, and training schedule. To avoid disadvantaging a heuristic ranking through an unsuitable pruning rule, we separately compare Top-K and Shift-Gsample for each ranking in Tab. 8 and report its better-performing variant in the main comparison.

B Training Details

B.1 Hyperparameters

To develop the rater, we first warm up the T2I proxy model using a fixed learning rate. Next, as described in Algorithm 1, the rater assigns a weight score to each sample, which is then used to reweight the training loss. Validation loss is employed as feedback to iteratively refine the rater through meta-learning.

The rating algorithm is optimized with AdamW with a learning rate of 1×10^{-4} and a batch size of 64.

Given the large-scale training set and the data efficiency (Dai et al., 2025) of modern generative models, we train STAR models from scratch for three epochs and fine-tune FLUX-mini with LoRA for one epoch unless otherwise stated. For from-scratch training, we use a total batch size of 512 with AdamW at a learning rate of 1×10^{-4}. For LoRA fine-tuning, we use rank $r = 16$ with a hidden size of 3072 and apply LoRA to all attention matrices, also using AdamW with the same learning rate.



Figure 7: Text–image pairs from different dataset domains.

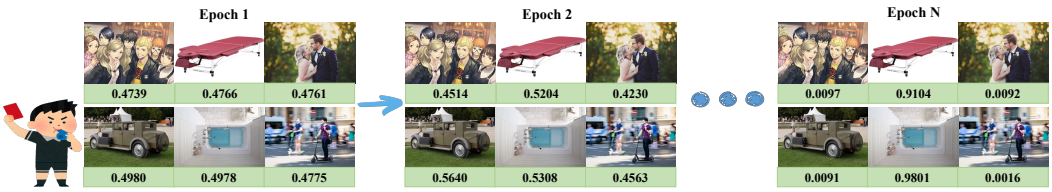


Figure 8: Changes in image ratings during training. The rater progressively refines its estimate of each sample's influence.

Shift-Gsample parameter protocol. The visually plain head region in Figure 4 of the main paper, together with the contiguous-range results in Table 4, motivates the fixed cutoff $n = 25$. We select the sampling center c and standard deviation σ using the MJHQ-30K ablation reported in Table 3(c) of the main paper. After selecting $n = 25$, $c = 50$, and $\sigma = 6$, we fix this configuration for GenEval, all other source datasets, and all downstream model settings without further tuning.

Table 5: Training results for random sampling, TASTE selection at different retention ratios, and full-data training on LAION. For each epoch, the best value is in **bold** and the second-best is underlined.

Epoch	20% Data					50% Data					Full Train		
	Random		Ours		Time	Random		Ours		Time	FID↓	CLIP↑	Time
	FID↓	CLIP↑	FID↓	CLIP↑	(h)	FID↓	CLIP↑	FID↓	CLIP↑	(h)			(h)
1	24.98	0.210	23.16	0.213	4.36	20.75	0.219	**18.49**	0.223	11.53	19.16	**0.225**	23.06
3	20.69	0.222	18.22	0.228	13.08	19.70	0.222	**16.20**	0.232	34.60	17.48	**0.234**	65.34
5	19.40	0.225	17.35	0.231	21.86	18.70	0.223	16.72	**0.234**	57.66	**16.50**	**0.234**	115.32
7	18.75	0.229	16.51	0.233	30.52	18.89	0.223	**16.02**	**0.235**	80.71	17.09	**0.235**	161.42
9	18.62	0.230	16.26	**0.235**	39.24	18.87	0.223	16.77	**0.235**	103.77	**15.82**	**0.235**	207.54

Table 6: Performance on LAION at different retention ratios. Avg. is computed as CLIP divided by FID. TASTE consistently outperforms random sampling, while the 50% subset remains comparable to full-data training with substantially less training time.

Selection %	Ours			Random			Time (h)
	FID↓	CLIP↑	Avg.↑	FID↓	CLIP↑	Avg.↑	
10%	21.54	0.219	0.010	22.15	0.217	0.010	6.54
20%	18.22	0.228	0.013	20.69	0.222	0.011	13.08
30%	17.80	0.227	0.013	20.03	0.223	0.011	19.59
40%	17.43	0.229	0.013	19.82	0.223	0.011	26.13
50%	**16.20**	0.233	**0.014**	19.70	0.222	0.011	34.60
100%	17.48	**0.234**	0.013	**17.48**	**0.234**	**0.013**	65.34

B.2 Rating Procedure and Score Dynamics

Our rater is updated in a data-driven manner by alternating between updating the proxy model and refining the rater over a large number of data samples. Algorithm 1 in the main paper shows the complete procedure, with our T2I-specific designs built upon the base meta-learning framework. With the trained rater, we apply it to the whole training dataset to assign an influence score to each data sample. As illustrated in Fig. 8, the scoring over the training set progressively differentiates according to each sample's intrinsic characteristics within the training dynamics. Across epochs, every sample's influence score evolves and gradually converges toward a stable estimate. This score reflects the estimated influence of each sample on model optimization: samples with higher scores typically correspond to lower training loss and faster convergence behavior. However, faster convergence does not necessarily correlate with better final performance. Therefore, an additional pruning strategy is required to select the samples that most effectively improve model generalization. The top score range mainly contains *plain* samples, the middle-to-late region contains *informative* samples, and the tail region contains *chaotic* samples. The fixed Shift-Gsample configuration improves downstream performance across the evaluated source datasets, suggesting that the pruning strategy is applicable to different data domains. TASTE-selected data aligns with human intuition, filtering out most plain and chaotic samples.

C Detailed Experimental Results

C.1 Quantitative Results

Reuse across downstream models. Tab. 7 reports the absolute results behind the relative improvements in Tab. 2(a) of the main paper. For a fixed source dataset, we train the rater and construct the selected subset once, then reuse that subset without architecture-specific reselection. The results show consistent improvements over random sampling for the evaluated STAR model scales and the FLUX-mini LoRA setting.

Table 7: Absolute Random and TASTE results underlying Tab. 2(a) of the main paper. The same selected subset is reused across the evaluated downstream model scales and architectures. STAR models are trained from scratch, while FLUX-mini-3B is fine-tuned with LoRA.

Downstream Model	Training	#Images	FID↓		CLIP↑	
			Random	Ours	Random	Ours
STAR-40M	From scratch	6M	28.66	**26.74**	0.201	**0.206**
STAR-0.3B	From scratch	3M	22.15	**21.54**	0.217	**0.219**
STAR-0.3B	From scratch	6M	20.69	**18.22**	0.222	**0.228**
STAR-0.3B	From scratch	15M	19.70	**16.20**	0.222	**0.233**
STAR-0.9B	From scratch	6M	18.14	**16.71**	0.233	**0.236**
FLUX-mini-3B	LoRA	6M	20.41	**19.98**	0.236	**0.237**

Table 8: Shift-Gsample versus Top-K for heuristic rankings on the 6M dataset after three epochs.

Ranking	Shift-Gsample		Top-K	
	FID↓	CLIP↑	FID↓	CLIP↑
Aesthetic	21.60	0.218	**17.48**	**0.228**
Clarity	21.90	0.218	**19.57**	**0.222**
Frequency	**19.41**	**0.222**	21.67	0.217
Edge-density	**21.14**	0.219	22.30	**0.221**

Pruning strategies for heuristic baselines. To ensure fair comparisons, we select the pruning rule separately for each heuristic ranking using a 6M-data validation study. As shown in Tab. 8, Aesthetic and Clarity perform better with Top-K, Frequency favors Shift-Gsample, and Edge-density obtains a better FID with Shift-Gsample while retaining a comparable CLIP score. We therefore report each heuristic baseline with its better pruning strategy in Tab. 1 of the main paper.

Stability on Different Epochs. As shown in Tab. 5, we evaluate the effectiveness of our method under a multi-epoch setting. With our TASTE-selected data training, each epoch achieves better results regardless of the training data scale. This saves training time by achieving superior training results within a given time period, enhancing efficiency in text-to-image model training.

Effectiveness under Different Data Selection Ratios. We evaluate the effectiveness of our method by selecting 10%–50% of the LAION full dataset (∼30M) and training for 3 epochs. The quantitative results are reported in Tab. 6. We define the average score as CLIP score divided by FID. Notably, our TASTE-selected 50% subset remains comparable to full-data training while requiring substantially less training time. This is because we filter out redundant and low-quality samples, which may appear risky as they could potentially disturb the gradient updates and hinder the model from reaching the optimal solution. Our experiments also show that models trained on TASTE-selected data consistently outperform those trained with randomly sampled subsets across all training stages. Moreover, we observe that the average score of random sampling stops improving even as more data are added. We attribute it to the inclusion of redundant and harmful samples, which can have no impact or degrade overall performance in models. In contrast, our selection strategy focuses on informative data, allowing the model to continue benefiting from meaningful training data samples.

Robustness across evaluation settings. The main-paper Tab. 1 evaluates the selected LAION subsets using complementary metrics on MJHQ-30K and GenEval. The absolute results in Tab. 7 further show improvements over random sampling across the evaluated STAR model scales and FLUX-mini. Main-paper Tab. 2(b) extends the comparison to HPDv3 and Flux-reason at two retention ratios. Together with the retention-ratio and training-dynamics results in Tabs. 6 and 5, these experiments show that the observed trends persist across evaluation metrics.

C.2 Qualitative Results

Our method applies to datasets from different domains. We use a web-crawled dataset (*i.e.*, LAION) and a synthetic dataset (*i.e.*, Flux-reason) to produce the qualitative results. Training on a TASTE-

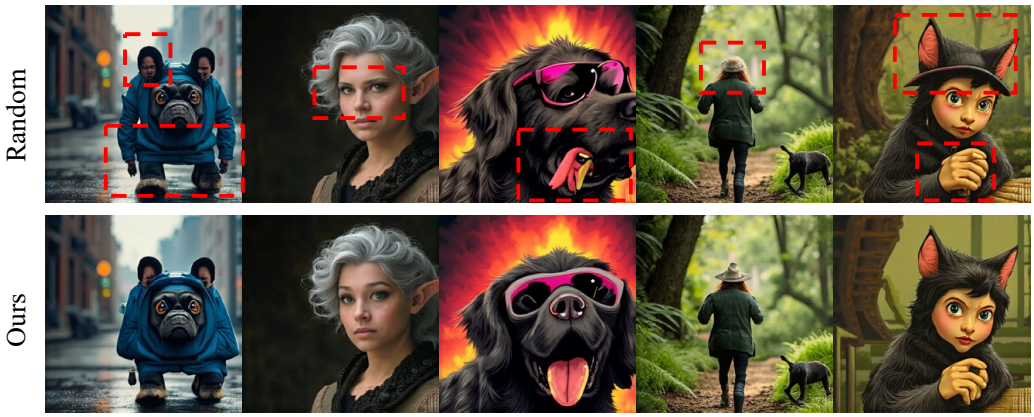


Figure 9: Qualitative comparison for FLUX-mini LoRA fine-tuning. Models are trained with the same data budget (∼78k iterations). Compared with random sampling, the subset selected by TASTE produces more coherent and semantically meaningful details while preserving the overall composition.

selected 50% subset remains comparable to full-data training and substantially outperforms random sampling. We further visualize the generated images from models trained with the same data volume. The FLUX-mini LoRA comparison is shown in Fig. 9. Results on GenEval datasets are shown in Fig. 10, and results on MJHQ-30K datasets are shown in Fig. 11.

C.3 Prompt Lists

We provide the prompts corresponding to the qualitative results presented in the paper.

Fig. 9

1. Design a superrealistic digital painting of a French Bulldog in a rainy urban scene, wearing a rain jacket. Use dim and moody lighting with street lights and reflections on the pavement to create the atmosphere. Make the rain jacket bright blue and the background dark blue and grey to contrast. Urban environment.

2. Halfelf with curly gray hair, big and pointy ears, nice, smiling, wearing medieval clothes, D&D character, fantasy world character, with background, photo taken by ARRI, photo taken by canon, photo taken by fuji, photo taken by kodak, incredibly detailed, professional lighting, film lighting, lightroom, cinematography, hyper realism, cinematic, film quality.

3. Retrowave labrador retriever wearing sunglasses with tongue out, lots of vivid color.

4. Yoda walking a Great Pyreneese dog in the jungle, 32k, rtx on.

5. Vintage anthropomorphic chonky mother cat, no children, in park, by Otto Dix, background.

Fig. 10

1. A photo of a bicycle

2. A photo of a suitcase

3. A photo of two backpacks

4. A photo of three pizzas, exclude four pizzas

5. A photo of a red apple

6. A photo of a red car

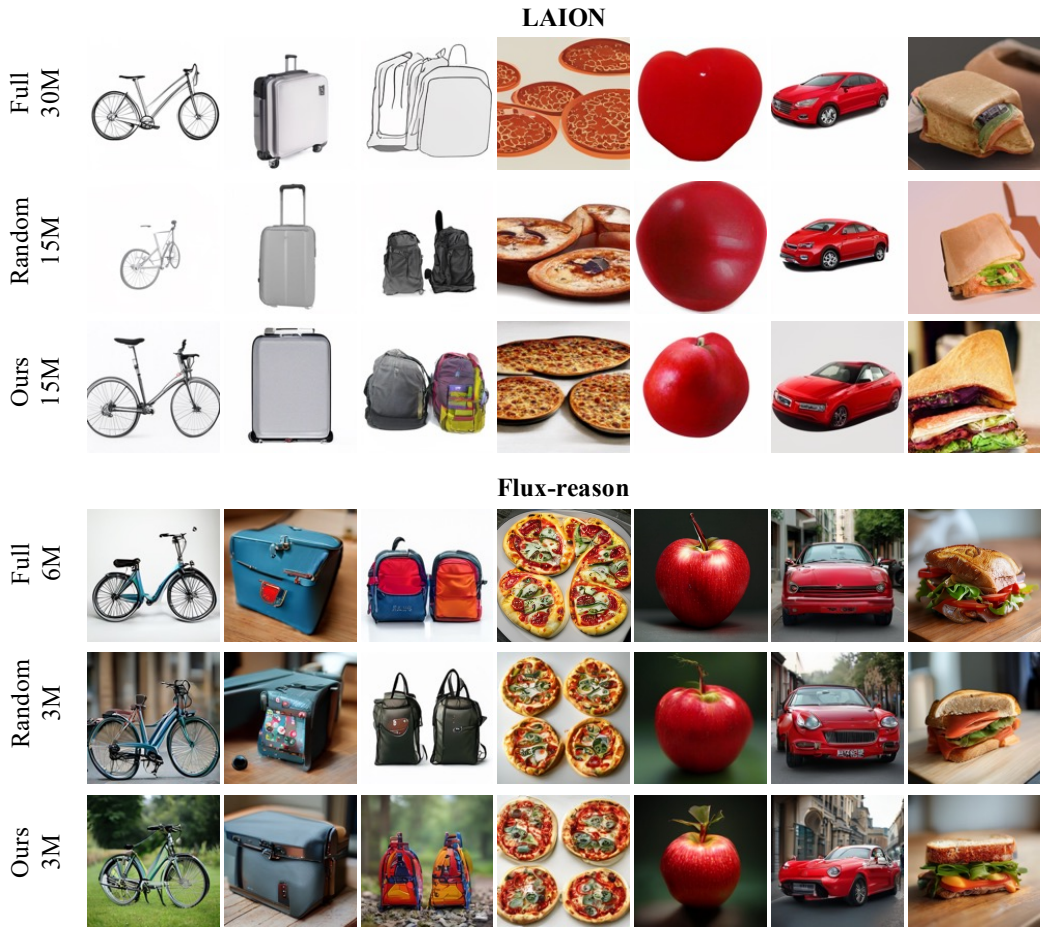


Figure 10: Qualitative comparison on GenEval under matched training iterations (∼30k) and batch size; all models therefore process the same number of training examples.

7. A photo of a backpack right of a sandwich

Fig. 11

1. Impressionistic oil painting in the style of Cuno Amiet, farmstead on rolling hills with wildflowers and a large oak tree.

2. Un piatto di carbonara in una trattoria italiana.

3. Photorealism of a spring morning, a pink flower blooms, light blue background, sunshine, super fine detail, ultra high definition image quality, 8K.

4. A Norfolk Island Pine in a pot, showcasing its soft, needlelike foliage.

5. Delicious orange chicken recipe served over steamed white rice, featuring bite-sized pieces of chicken coated in a tangy and glossy orange sauce, garnished with thinly sliced green onions and a sprinkle of sesame seeds on top.

6. Beautiful landscape with sunrise, nature, peaceful, sunlight, bright, serene, beautiful sky, 32k, relaxation, sublime, calming.

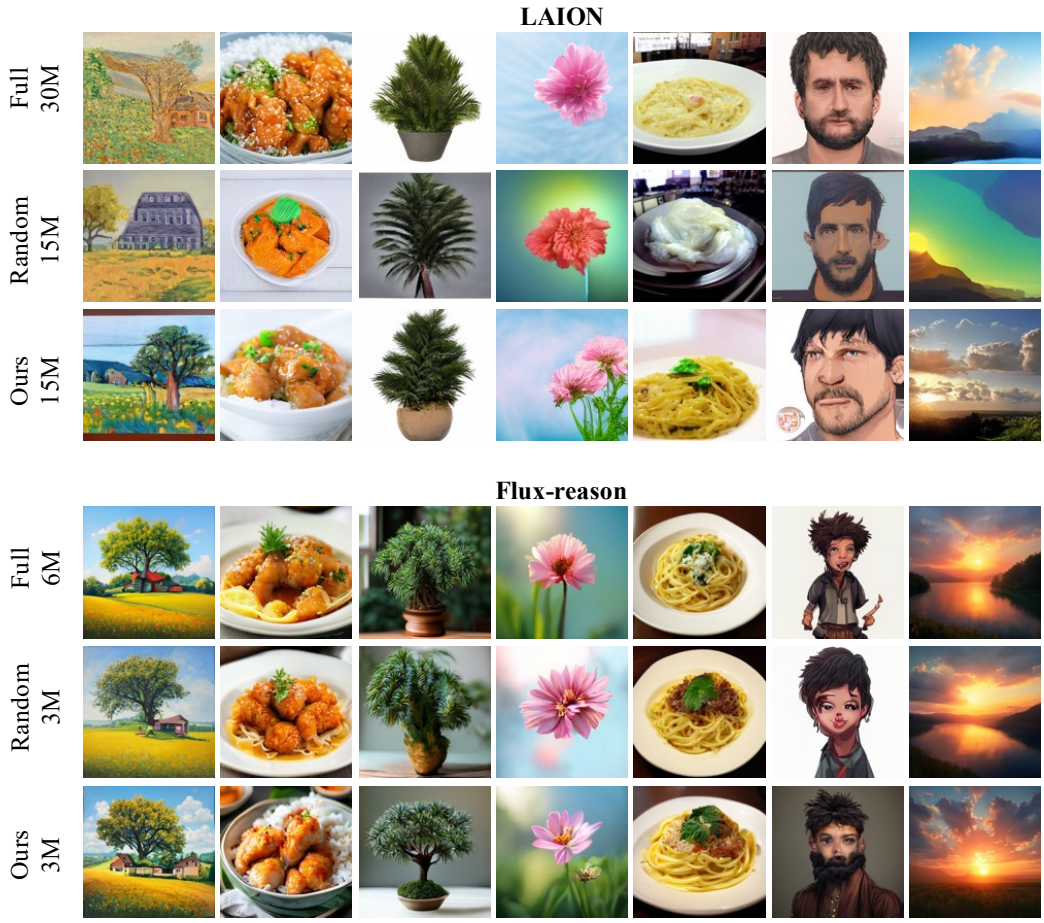


Figure 11: Qualitative comparison on MJHQ-30K under matched training iterations (∼30k) and batch size; all models therefore process the same number of training examples.

7. Character design, a Mediterranean guy with slightly curly medium dark-brown long hair, thin Greek mustache, nice lips, slightly almond-shaped eyes.

D Limitations and Future Work

Our approach focuses on selecting training-beneficial data but relies on uniformly shuffling the chosen samples during training. While this strategy is simple and effective, we do not exploit potential benefits from curriculum learning. Presenting data in an "easy-to-hard" sequence may further stabilize optimization and improve convergence, especially for large-scale datasets that can only be traversed once. Designing principled curricula or adaptive sampling strategies with data ordering therefore remains an important direction for future research.

Across the evaluated downstream settings, the selected subsets transfer from the autoregressive proxy to STAR models of different scales and to FLUX-mini with LoRA. However, substantially larger target models may exhibit training dynamics that differ from those of the proxy, potentially reducing the reliability of the learned ranking. Due to limited computational resources, we do not evaluate the selected subsets through from-scratch training of substantially larger T2I models. Developing more robust cross-scale rating strategies and validating them through large-model from-scratch training remain important directions for future work.