FINE-T2I: AN OPEN, LARGE-SCALE, AND DIVERSE DATASET FOR HIGH-QUALITY T2I FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality and open datasets remain a major bottleneck for text-to-image (T2I) fine-tuning. Despite rapid progress in model architectures and training pipelines, most publicly available fine-tuning datasets suffer from low resolution, poor text–image alignment, or limited diversity, resulting in a clear data gap for open text-to-image research. In this work, we present Fine-T2I, a large-scale, high-quality, and fully open dataset for open T2I fine-tuning. Fine-T2I spans 10 task combinations, 32 prompt categories, 11 visual styles, and 5 prompt templates, and combines synthetic images generated by strong models with carefully curated real images from professional photographers. Unlike previous datasets that simply present a set of samples, Fine-T2I is built with: **1)** systematic dataset processing pipeline, **2)** principled distribution design, **3)** multi-granularity prompt design, **4)** dynamic aspect-ratio & resolution strategy, and **5)** advanced reasoning-VLM-as-a-judge protocol. Furthermore, all samples are rigorously filtered for text–image alignment, visual fidelity, and prompt quality, with over 95% of initial candidates removed. The final dataset contains over 6 million text–image pairs, around 2 TB on disk, approaching the scale of pretraining datasets while maintaining fine-tuning-level quality. Across the open diffusion and autoregressive models we study, fine-tuning on Fine-T2I improves generation quality and instruction adherence, as validated by human evaluations and visual comparisons, with gains that depend on the base model's starting capability. Our analysis further shows that data aligned with a benchmark's distribution can inflate automatic scores without improving human-judged quality, underscoring the value of preference-oriented curation. We release Fine-T2I under an open license to help close the data gap in the open community.

1 INTRODUCTION

Text-to-image (T2I) generation has advanced rapidly over the past two years, enabling the synthesis of highly aesthetic images while faithfully adhering to diverse and complex instructions. This progress has been driven by a new wave of frontier generative models, including GPT Image 1.5, Nano Banana Pro Team et al. (2025), Seedream 4.0 Seedream et al. (2025), and Qwen-Image Wu et al. (2025a), among others. Such achievements largely stem from two complementary factors: **model-centric** advances in architectures Wu et al. (2025a); Gong et al. (2025), training objectives Liu et al. (2025a;b), and inference techniques Ma et al. (2025a); and **data-centric** progress in dataset scale, curation, and quality. Together, they push the frontier of high-fidelity image generation.

However, these two factors do not propagate equally across the open community. Model-centric advances, such as architectures, training objectives, and inference techniques, are routinely shared through publications or blogs and adopted broadly. In contrast, data-centric resources, especially carefully curated and instruction-aligned fine-tuning data, remain largely private and inaccessible. As a result, SOTA models have become increasingly concentrated in top industry groups. Even when the open community matches the model training recipes and has sufficient compute, the lack of comparable high-quality data can lead to a persistent and potentially growing gap. Indeed, the text-to-image leaderboards are always dominated by closed frontier products.

Similar to the open dissemination of model-centric techniques, the community has a strong interest in open-sourcing *large-scale, high-quality* data for T2I training and alignment. Yet this is difficult for three reasons. First, truly high-quality fine-tuning images are extremely expensive (often $10+ per

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.

Dataset	High Resolution?	High Quality?	Designed Prompts?	Diverse Resolutions?	Distribution Analysis?	Large Scale?
JourneyDB Sun et al. (2023)	✓	✓	✗	✓	✗	✓
Pick-a-Pic Kirstain et al. (2023)	✗	✓	✗	✗	✗	✗
Pick-a-Pic-v2 Kirstain et al. (2023)	✗	✓	✗	✗	✗	✓
T2I-2M jackyhate (2024)	✗	✗	✗	✗	✗	✓
LAION-Art Schuhmann et al. (2022)	✗	✗	✗	✓	✗	✓
LAION-Aesthetic Schuhmann et al. (2022)	✗	✗	✗	✓	✗	✓
Blip3o-60k Chen et al. (2025)	✗	✓	✗	✗	✗	✗
Fine-T2I (ours)	✓	✓	✓	✓	✓	✓

Table 1: We compare our Fine-T2I with open-sourced fine-tuning datasets. High-resolution indicates the resolution of most samples is greater than 1K. More details can be found in Sec. 3. See Fig. 12 for a qualitative visual comparison illustrating the data quality.

image). Second, such images are typically restricted by licenses and cannot be legally redistributed. Third, how to build a promising high-quality dataset is unclear and still underexplored. Consequently, while many companies open-source production-grade T2I models, high-quality fine-tuning datasets are rarely released, leaving the open community to rely on datasets that are comparatively small, noisy, or distribution-specific (Table 1). Although these datasets laid the groundwork for the field, a clear gap in scale and quality remains for modern T2I fine-tuning.

To alleviate this bottleneck, we introduce Fine-T2I, a large-scale, high-quality, and fully open dataset for text-to-image fine-tuning. The Fine-T2I dataset combines *(a)* synthetic data generated by strong diffusion models with *(b)* curated real images licensed and shared by photographers. For the synthetic portion, rather than collecting existing text-image pairs, we build the dataset from scratch: we carefully design a diverse, weighted distribution over styles, categories, tasks, and prompt templates (Table 1), and adopt a multi-granularity prompt design, enhancing each original prompt into a longer, more detailed counterpart with a fine-tuned enhancer Wang et al. (2025b), together with a dynamic aspect-ratio and resolution strategy instead of fixed square images. For the curated real-image subset, we use a strong VLM to generate both short and long prompts. To ensure high quality, we apply a detailed filtering pipeline (de-duplication, text–image alignment checks, safety filtering, and aesthetic assessment with an advanced reasoning-VLM-as-a-judge protocol), removing over 95% of the initially collected data. The resulting dataset contains roughly 6 million examples and about 2 TB on disk. Fig. 1 shows examples from our Fine-T2I dataset. Unlike other open datasets that often assume fixed resolutions, fixed prompt templates, or simplified settings, Fine-T2I targets a more practical goal: providing a strong, openly available fine-tuning resource that reflects real-world generation requests for the open community.

We validate Fine-T2I across diverse T2I architectures, including autoregressive, latent-diffusion, and pixel-space diffusion models, and observe consistent improvements under human studies, qualitative comparisons, and automatic metrics. We hope Fine-T2I and its construction pipeline serve as a practical foundation for open image generation.

2 Related Work

Text-to-Image Generation Currently, most text-to-image generation models rely primarily on two architectural designs: autoregressive models Sun et al. (2024); Wang et al. (2024); Ma et al. (2025b); Tian et al. (2024); Li et al. (2024) and diffusion models Ho et al. (2020); Rombach et al. (2022); Lipman et al. (2022); Dai et al. (2023); Wu et al. (2025a). Regardless of the differences in architectural designs, these models typically follow a progressive training strategy. First, we begin with large-scale pretraining on massive text-image-pair datasets, like LAION-5B Schuhmann et al. (2022) and OpenImages Kuznetsova et al. (2020), to learn broad world knowledge. Next, models are fine-tuned on curated, high-aesthetic datasets, shifting the distribution to enhanced visual fidelity and detailed generation. Recent methods also incorporate Reinforcement Learning from Human Feedback (RLHF) such as Direct Preference Optimization (DPO) Rafailov et al. (2023); Esser et al. (2024); Wang et al. (2024) or Reinforcement Learning (RL) Liu et al. (2025a); Seedream et al. (2025); Wu et al. (2025b); Wang et al. (2025a). Crucially, this alignment is not a one-time process: recent multi-turn strategies iterate between human-feedback alignment and high-quality fine-tuning Deng et al. (2025), further underscoring the importance of SFT data.

Fine-tuning Datasets Table 1 provides a detailed summary for popular fine-tuning datasets. While these datasets have laid the groundwork for open image generation research, they still exhibit clear

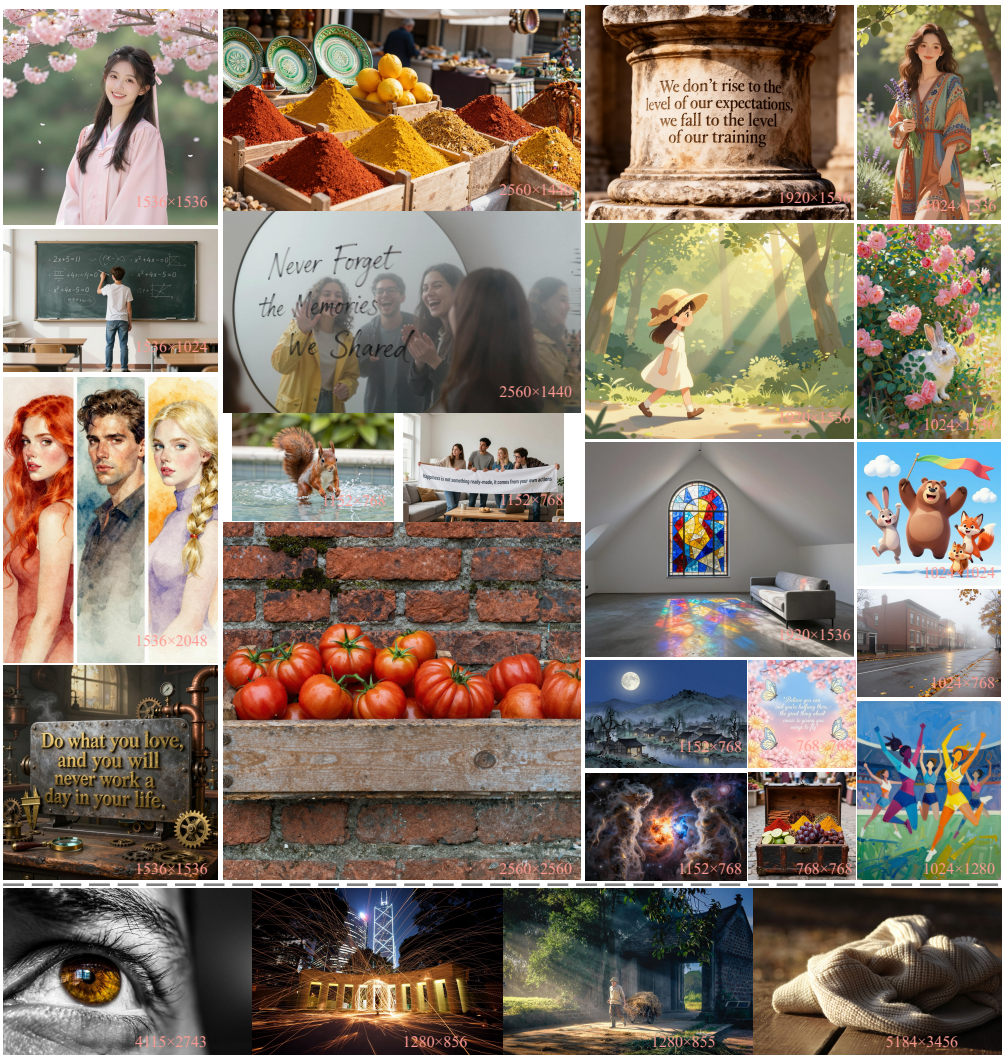


Figure 1: Examples of our introduced ***Fine-T2I*** dataset samples, which include diverse resolutions, styles, categories, tasks, *etc*. Please check the supplementary for details. Images above the dashed line are our synthetic samples, and those below the dashed line are our curated real images.

gaps relative to modern production-grade models. First, many samples from these datasets are low-resolution (often $\leq 1024 \times 1024$), whereas recent models (*e.g.*, Qwen-Image Wu et al. (2025a), SeedDream Gao et al. (2025); Seedream et al. (2025), and Nano Banana Pro Comanici et al. (2025)) natively generate images at substantially higher resolutions. Second, even after applying aesthetic filters like LAION-Art and LAION-Aesthetic, the resulting data often falls short of the visual fidelity and aesthetic (see Fig. 12) expected by today's models. Importantly, while often overlooked, existing open datasets rarely provide careful distribution analysis and organization, despite growing evidence that such curation is critical for performance Wu et al. (2025a); Seedream et al. (2025); Cai et al. (2025). These limitations motivate the need for a modern, high-quality fine-tuning dataset.

3 Fine-T2I Dataset

We introduce **Fine-T2I**, a new fine-tuning dataset designed for modern high-quality image generation in the open community. In this section, we describe the full pipeline for building our dataset.

3.1 Generating Synthetic Set

The synthetic set constitutes the majority of Fine-T2I. We first construct controlled prompts with LLMs and apply rigorous filtering and de-duplication. Each retained prompt is then enhanced into

a detailed, more descriptive counterpart using a fine-tuned prompt enhancer Wang et al. (2025b). Using both the original and enhanced prompts, we generate images with SOTA diffusion models Cai et al. (2025); Labs (2025) under two generation settings: (i) square aspect ratio and (ii) randomized resolutions. This yields four synthetic sub-datasets in total. Finally, we perform additional filtering to remove low-quality and poorly aligned samples, resulting in 6,145,693 high-quality samples.

3.1.1 Designing a Diverse Prompt Distribution

Without careful design, synthetic prompt distributions skew toward dominant scenarios, encouraging overfitting and underrepresenting long-tail cases. To obtain a distribution that is both diverse and representative of real usage, we follow the dataset analyses reported in recent public reports (e.g., Qwen-Image Wu et al. (2025a) and Seedream Seedream et al. (2025); Gong et al. (2025)) and complement them with trends reflected by the Image Arena leaderboard. We generate synthetic prompts using the LLaMA3 instruction model Grattafiori et al. (2024). Concretely, we design a *weighted* taxonomy of high-level categories, including *Nature*, *Design*, *People*, *Text rendering*, and a set of *rare cases*. We further decompose each category into fine-grained sub-categories to improve coverage and controllability. We then synthesize *controlled* prompts by composing the sampled category with several predefined attributes, including weighted style, prompt length, prompt templates (see supplementary Sec. A), and task type. Categories and styles are sampled by calibrated weights approximating real-world usage: we emphasize frequently requested content (nature, people) and broadly useful styles (photorealistic), while keeping non-trivial mass on long-tail cases. The remaining attributes are sampled uniformly at random to increase variability. To further reduce templated outputs, we set a higher sampling temperature 1.4 and a lower nucleus threshold top-$p = 0.8$. We additionally filter out trivial generations by discarding prompts shorter than five words. In total, this yields 44,800,567 controlled prompts. We discuss practical issues encountered during prompt generation in Sec. B.

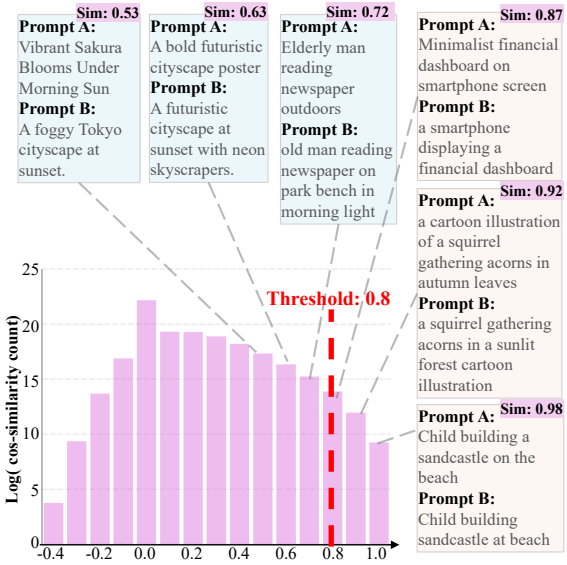


Figure 2: Semantic cosine-similarities distribution in a random prompt subset. We set deduplication threshold to 0.8.

3.1.2 Semantic Prompt De-duplication

In practice, prompt de-duplication Wu et al. (2025a); Cai et al. (2025); Cao et al. (2025) is commonly implemented either via (i) MinHash-LSH over token shingles (n-grams) Wu et al. (2025a) or (ii) embedding-based semantic filtering. Although fast and scalable, MinHash is largely lexical and may remove prompts that share phrasing but differ in meaning. Since our prompts are LLM-generated and often vary in meaning with critical details under similar syntactic patterns, we adopt *semantic* de-duplication.

We encode each prompt with the all-MiniLM-L6-v2 sentence encoder Reimers & Gurevych (2019); Wang et al. (2020) into a 384-dimensional embedding. Computing full pairwise similarities is infeasible at our scale, so we employ a two-stage hierarchical strategy. We treat two prompts as duplicates if their cosine similarity exceeds 0.8; see Fig. 2 for illustration. After semantic de-duplication, we retain 5,555,147 diverse prompts.

3.1.3 Prompt Filtering

Safety Guard To ensure the resulting prompts are safe and compliant, we apply content filtering with LLaMA-Guard-3-8B Inan et al. (2023). We remove prompts flagged as *violent crimes*, *child*

sexual exploitation, *privacy*, and other unacceptable content. In addition, we filter by length and drop prompts longer than 150 words. Overall, this step provides 5,497,062 safety-checked prompts.

Attributes check In addition to the safety guard, we further filter prompts by checking whether each generated prompt aligns with its intended control attributes. Here, we focus on *style* and *category* consistency. We use Qwen3-VL-8B-Instruct Bai et al. (2025a) as an attribute verifier; prompts that fail the alignment check are discarded. After attribute-alignment filtering, we obtain a final set of 5,158,969 prompts for subsequent image generation. See Sec. B for more discussions.

3.1.4 Multi-Granularity Prompt Design

Rewriting user inputs into more detailed instructions is increasingly used at inference to improve quality and alignment Ma et al. (2025b); Wu et al. (2025a); Comanici et al. (2025); we argue it is also valuable *during training*. Motivated by this, we explicitly introduce *Multi-Granularity Prompt Design* in Fine-T2I. For each filtered prompt, we apply a fine-tuned prompt enhancer Wang et al. (2025b) to rewrite it into a longer, more descriptive counterpart while maintaining the original intent. The enhanced prompts add visual attributes and reduce ambiguity; we keep both forms so that models see diverse input granularities.

3.1.5 Dynamic Aspect Ratio and Resolution Strategy for Image Generation

We generate synthetic images from both prompt sets in Fine-T2I: the *original* generated prompts and their *enhanced* (rewritten) counterparts. For each variant, we generate images under two complementary settings. The first uses *randomized resolutions and aspect ratios* to better reflect real-world generation requests and to encourage robustness beyond a single canonical format. Please check Tab. 8 for detailed resolutions. The second restricts *square* images with multiple square sizes, since many existing pipelines and benchmarks still primarily consider square inputs.

For each prompt, we prioritize quality via a generate-and-select strategy. We sample 1-3 candidate images using two strong open-source generators, Z-Image Cai et al. (2025) and FLUX2 Labs (2025), and keep the candidate with the highest Aesthetic Predictor V2.5 score[1]. Compared with the FLUX Labs et al. (2025); Labs (2024), GPT-4o, and MidJourney generations used in T2I-2M, BLIP3-o, and JourneyDB, our generators provide substantially higher quality and aesthetics, which is critical for a high-quality fine-tuning set. Combining the two prompt variants with the two resolution settings yields four synthetic subsets: *Enhanced+Square*, *Enhanced+Random*, *Original+Square*, and *Original+Random*. Since FLUX2 is substantially more expensive to run, more images in Fine-T2I are generated with Z-Image. Please check Fig. 1 for generated examples.

3.1.6 Reasoning-VLM-as-a-Judge for Text-Image Pair Filtering

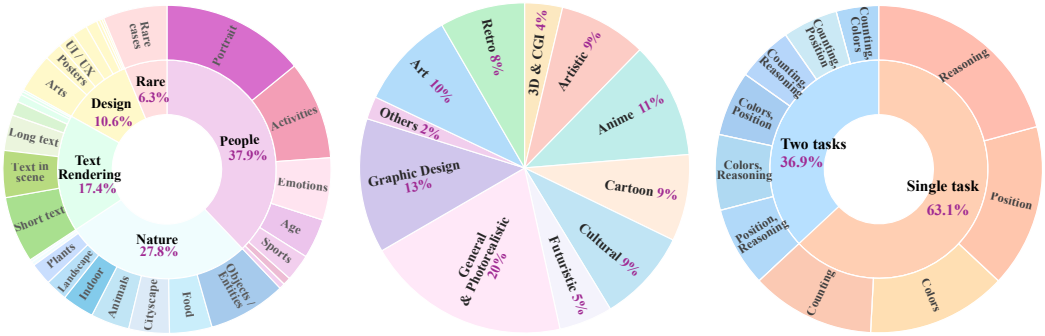
Although we use strong generators, the generated images are not always reliable. In practice, failures mainly arise from two reasons: (i) insufficient visual quality and aesthetics, and (ii) imperfect text-image alignment. We therefore apply a two-stage filtering pipeline for these issues.

We first filter images using Aesthetic Predictor V2.5 that focuses on image aesthetics. We keep an image only if its score > 5.5, which indicates very high visual quality. Filtering text-image alignment is substantially more challenging. Scalar human-preference scores such as HPSv2 (Wu et al., 2023) and HPSv3 (Ma et al., 2025c) are widely used, but recent studies show that they compress multiple quality dimensions into a single value and can overlook fine-grained errors while favoring aesthetic appeal, and are prone to reward hacking that yields visible artifacts (Kao et al., 2026; Hong et al., 2026; Ma et al., 2025c). We therefore avoid using such scores as our filtering criterion here. Instead, we verify alignment and check artifacts with a reasoning ("thinking") VLM. Guided by a carefully designed system prompt that enumerates the alignment and artifact criteria (Fig. 11), the VLM accurately flags prompt non-compliance and typical failure modes, and we discard any sample that violates

Subset	PE-AR	PE-AS	PO-AR	PO-AS
Samples	1,615,592	1,538,253	1,686,498	1,305,350
Storage	476G	517G	479G	436G

Table 2: Number of samples for the synthetic set after filtering. "PE" indicates enhanced prompts, "PO" denotes original prompts, "AR" stands for random aspect ratios, and "AS" means square aspect ratios.

[1] https://github.com/discus0434/aesthetic-predictor-v2-5



(a) Categories Analysis (b) Styles Analysis (c) Tasks Analysis

Figure 3: Analysis for our synthetic set, we provide the distribution for sample categories, sample styles, and the related tasks. Please check Sec. A in the supplementary for details.

any criterion. Despite its higher cost, this step is crucial for meeting the fine-tuning quality bar: on average it removes about 70% of text-image pairs, yielding the final counts in Table 2.

3.2 Curating Real-Image Set

Fine-T2I also includes a curated real-image set. Curating "high-quality" real data is ambiguous: aesthetic quality is hard to formalize, and automatic scorers behave unpredictably on professional photography (*e.g.*, subtle composition and lighting are not captured by a single score; see Sec. B). We therefore start from creator-driven platforms: *Pexels*, *Pixabay*, and *Unsplash-Lite*, where photographers upload and curate under open terms, giving a strong aesthetic prior with broad photographic coverage.

Data Source	Downloaded	After Filter	Kept ratio	Storage
Pexels	233,342	117,389	50.3%	192.6 GB
Pixabay	163,524	32,654	20.0%	9.7 GB
Unsplash_lite	24,997	18,381	73.6%	56.2 GB

Table 3: Number of samples for curated real-image set. We curated about 168k extremely high-quality samples from open-sourced platforms for this set in Fine-T2I.

We then filter with Aesthetic Predictor V2.5, keeping only scores ≥ 6.5 (a stricter bar for this subset), and discard extremely large images for training stability.

To make these images qualified for text-conditioned training, we generate captions with a fine-tuned[2] Qwen2.5-VL-7B model Bai et al. (2025b). Same as the synthetic set, we provide two prompts for each image: an initial prompt (short length) and a more detailed enhanced version generated by the same prompt enhancement procedure. This mirrors the way prompts are used in practice, from short user inputs to more explicit, descriptive instructions, and provides richer instructions for fine-tuning. After filtering and captioning steps, the real-image subset contains 168,424 images. See detailed statistics in Table 3.

4 Fine-T2I Dataset Specifications

Analysis on the attributes Fig. 3 summarizes the distributions over **categories**, **styles**, and **tasks**. As shown in Fig. 3a, the category mixture focuses on realistic, high-demand content: *People* (37.9%) and *Nature* (27.8%) form the majority, and *Rare cases* (6.3%) ensure non-trivial long-tail support. Fig. 3b indicates that styles are similarly balanced: *General & Photorealistic* is the largest portion (20%), complemented by diverse artistic and design-oriented styles, which helps the dataset span both photorealistic and stylized creation. Finally, Fig. 3c shows that most prompts are *single-task* (63.1%) with a sizable fraction of *two-task* compositions (36.9%); see Fig. 8 (supplementary) for the prompt-length distribution.

Although these empirical distributions deviate slightly from the initial sampling weights (due to non-uniform rejection during filtering), the final dataset still closely matches the intended real-world demand: it emphasizes common user scenarios while preserving long-tail and multi-attribute coverage, which is critical for training strong modern T2I models.

[2]https://huggingface.co/Ertugrul/Qwen2.5-VL-7B-Captioner-Relaxed

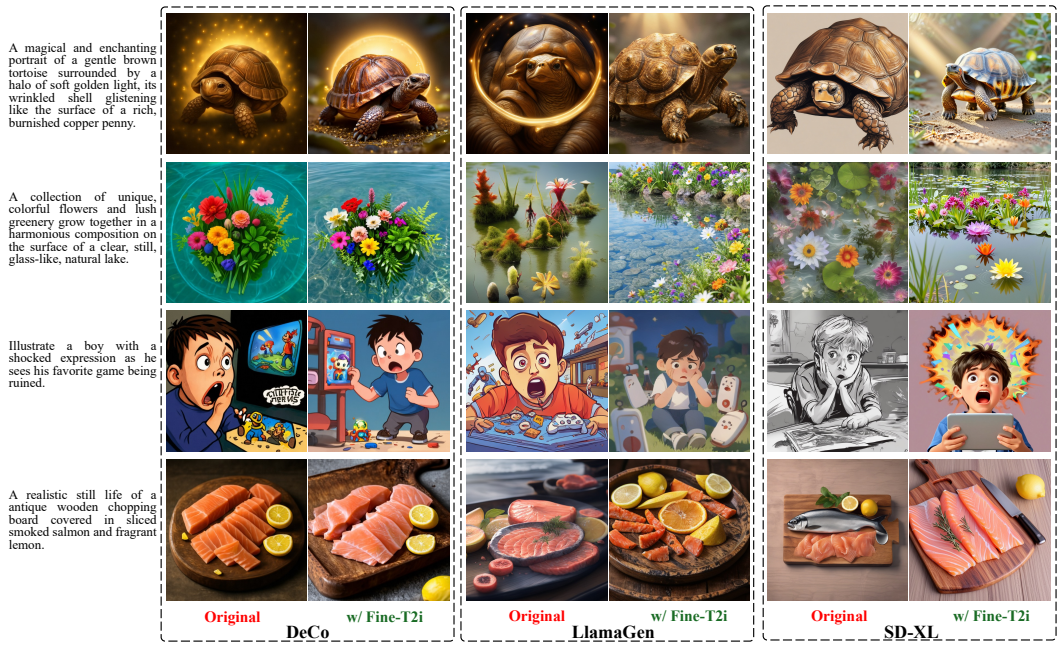


Figure 5: Visual comparison between original results and Fine-T2I fine-tuned results with different models. Clearly better generation quality can be observed when fine-tuning with Fine-T2I.

Analysis on the aesthetic quality We examine aesthetic quality with Aesthetic Predictor V2.5 scores for the synthetic and curated sets (Fig. 4). For this metric, scores above 5.5 are generally regarded as indicative of high aesthetic quality. First, the curated real-image subset concentrates at higher scores (majority in $[6.5, 7.0)$, 63.21%, with a non-trivial tail beyond 7.5), consistent with our creator-platform sourcing and stricter filtering. Second, the *synthetic subset* exhibits a broader score distribution. Most samples lie in the range of $[5.5, 6.5)$ (36.64% in $[5.5, 6.0)$ and 35.70% in $[6.0, 6.5)$), while a meaningful fraction reaches higher aesthetic ranges, with a small tail above 7.5. This reflects the inherent variability of generative models and our use of a permissive threshold to preserve scale and diversity. Overall, the two subsets are complementary. The curated real-image set provides a consistently high-aesthetic standard, while the synthetic set contributes large-scale, diverse guidance with controlled attributes.

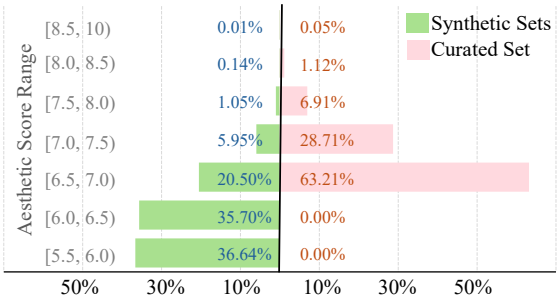


Figure 4: The aesthetic score distribution of our Fine-T2I.

5 Experiments with Fine-T2I

We next evaluate Fine-T2I for image generation, asking whether it consistently improves generation quality and instruction following across model families.

Experimental Setting We consider three representative T2I backbones spanning different generation technologies: latent-diffusion model SD-XL Podell et al. (2023), autoregressive model LlamaGen Sun et al. (2024), and the most recent pixel-space diffusion model DeCo Ma et al. (2026). We select these models because they are not well-trained on closed high-quality datasets like Qwen-Image or Nano-Banana, better demonstrating the performance gap. For each model, we start from the official publicly released pretrained checkpoint and perform continued fine-tuning on Fine-T2I. Fine-T2I includes four synthetic subsets (original/enhanced $\times$ square/random resolution) as well as a curated real-image set; unless otherwise specified, we fine-tune the models using a mixture of all sub-

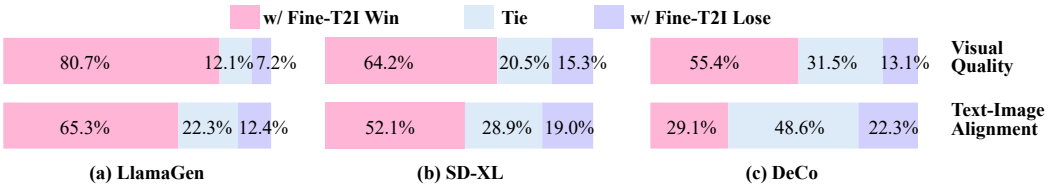


Figure 6: Human-evaluation results for LlamaGen, SD-XL, and DeCo, comparing generations produced with and without fine-tuning on our dataset.

sets, i.e., the released dataset as a whole; a controlled ablation of the individual pipeline components is left to future work. For SD-XL, we fine-tune with LoRA adapters to ensure a lightweight training setup, a batch size of 8 on each GPU and a learning rate of 1×10^{-4}. For LlamaGen and DeCo, we fine-tune the full model using batch size 24 and learning rate 3×10^{-5} and batch size 16 and learning rate 1×10^{-5}, respectively. We filter out samples with relatively long prompts, considering the limited token length supported by the three models. We fine-tune each model for about one epoch to avoid overfitting; tuning these settings further is not our focus.

Evaluation As noted in HunyuanImage Cao et al. (2025) and related work, commonly used automatic benchmarks such as GenEval Ghosh et al. (2023) and T2I-CompBench Huang et al. (2023) have limited coverage and can be misaligned with human preference, due to restricted prompt diversity. *This limitation of auto-evaluation is especially pronounced when the goal is to assess the quality of a fine-tuning dataset*, where improvements may manifest in subtle aspects such as aesthetics, style fidelity, and nuanced instruction following.

To better reflect real-world usage, we build an evaluation suite of 500 public prompts randomly sampled from the Artificial Analysis Image Arena leaderboard, which are widely used to compare leading T2I models. We run large-scale human preference evaluation on the generations, focusing on text-image alignment and overall visual quality. For completeness we also report GenEval (Table 4), though we caution against relying heavily on such automatic metrics.

Results analysis We primarily evaluate the improvements from fine-tuning on Fine-T2I based on human evaluation, since our goal is to assess Fine-T2I as a *fine-tuning dataset* and the improvements we target (overall aesthetics, realism, and instruction following) are not reliably captured by automatic benchmarks. As shown in Fig. 6, fine-tuning on Fine-T2I leads to clear preference gains on all three models. By fine-tuning LlamaGen on our dataset, we achieve an 80.7% win rate for visual quality compared to a counterpart without further fine-tuning, and a 65.3% win rate for text-image alignment. We also see substantive improvements for SD-XL and DeCo. These observations suggest that our Fine-T2I is able to boost different architectures and generation designs, providing consistent benefits across different model families. We also notice that our dataset boosts the visual quality the most. This might be due to our extremely high aesthetic quality.

Model	Single	Two	Count	Color	Position	Attribute	Overall
LlamaGen	0.71	0.34	0.21	0.58	0.07	0.04	0.32
- w/ Fine-T2I	0.91	0.44	0.13	0.68	0.12	0.09	0.39 (+0.07)
SD-XL	0.98	0.64	0.35	0.77	0.11	0.16	0.50
- w/ Fine-T2I	0.98	0.71	0.46	0.84	0.10	0.20	0.55 (+0.05)
DeCo	1.00	0.92	0.66	0.77	0.72	0.77	0.82
- w/ Fine-T2I	0.99	0.89	0.60	0.92	0.54	0.68	0.77 (-0.05)

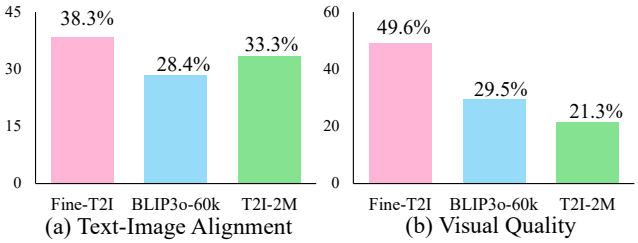
Table 4: GenEval results for LlamaGen, SD-XL, and DeCo. We re-evaluate all models instead of reported results.

The qualitative examples are presented in Fig. 5. After fine-tuning on Fine-T2I, generations look more natural and coherent, with much improved aesthetic quality and fewer distracting artifacts, while better reflecting the intended prompt content. *While we do not consider GenEval to be conclusive evidence,* the GenEval scores improve for LlamaGen and SD-XL, as reported in Table 4, offering an additional signal that Fine-T2I improves controllability under standard automatic protocols. For DeCo, the overall score slightly decreases, which we attribute to a benchmark-specific effect rather than a quality regression: DeCo is trained on the BLIP3o dataset (Ma et al., 2026; Chen et al., 2025), whose 60k-sample instruction-tuning set (BLIP3o-60k) includes a *Geneval* subset drawn from the GenEval prompt distribution, with the test set held out.[3] Its base GenEval score

[3]See the dataset card at `https://huggingface.co/datasets/BLIP3o/BLIP3o-60k`. DeCo reports an overall GenEval of 0.86, whereas we report 0.82 from a unified re-evaluation (Table 4).

is thus obtained partly in-distribution with the benchmark, a property shared by any model trained on this set, including BLIP3o itself, and fine-tuning on the different, human-preference-oriented distribution of Fine-T2I moves the model away from this regime. This reading is consistent with our human evaluation, where DeCo still improves after fine-tuning (Fig. 6).

Comparing with other SFT datasets Previous studies demonstrate the high quality of our Fine-T2I and the benefits of our dataset. Here, we compare our dataset with other datasets. We use LlamaGen as the backbone for simplicity. We fine-tune LlamaGen with the same settings on T2I-2M, BLIP3o-60k, and our Fine-T2I, respectively. We conduct a human evaluation (select the best from three generations) on the 500 prompts mentioned above, and report the win rate results in Fig. 7. The generated images can be found in Fig. 10 for visual comparison. As these datasets span orders of magnitude in size, we fine-tune on each at its natural scale for one epoch, reflecting realistic use rather than a matched-budget ablation; on both text alignment and visual quality, models fine-tuned on Fine-T2I are clearly preferred. Importantly, this gain is not merely an effect of scale: the far smaller BLIP3o-60k reaches the lowest training loss yet yields the weakest generations (Appendix C), indicating that quality rather than quantity drives the improvement, and delivering high quality at scale is precisely what Fine-T2I contributes. We further refer to Fig. 12 for a sample-level comparison of the datasets themselves.



(a) Text-Image Alignment (b) Visual Quality

Figure 7: Human evaluation results (win rate) comparing models fine-tuned on different datasets. The best generation result among the three is selected as the winning example.

Analysis: what makes fine-tuning data effective? Our results also caution about *how* improvement is measured. We find that data aligned with a benchmark's own distribution can raise that benchmark's score without a matching gain in human-preferred quality, and two observations from our experiments point to this. First, as discussed above, DeCo's high base GenEval is obtained partly in-distribution with the benchmark, since its BLIP3o-60k training data contains GenEval-distribution prompts; fine-tuning on Fine-T2I lowers GenEval yet still improves human preference. Second, among the datasets we compare (Fig. 7), fine-tuning on BLIP3o-60k yields the lowest training loss but the weakest generations (Appendix C): its structurally simple, benchmark-shaped samples are easy to fit but transfer poorly to real generation quality. Together, these indicate that optimizing benchmark-aligned data is not the same as optimizing quality. Fine-T2I instead targets human-preferred quality through its distribution design, aesthetic filtering, and reasoning-VLM verification (Sec. 3.1.6), which is why its benefits are clearest under human evaluation; we further discuss the gap between automatic aesthetic scores and human preference in Appendix B.

6 Conclusion

Instead of model-centric innovations, publicly available high-quality fine-tuning data remains a key bottleneck for high-quality text-to-image generation in the open community. To address this gap, we introduce Fine-T2I, a large-scale, fully open dataset that combines diverse synthetic text–image pairs with a rigorously curated real-image set. Fine-T2I is built with a systematic data design and advanced filtering pipeline that prioritizes diversity, alignment, and visual quality at scale. We show that fine-tuning on Fine-T2I consistently improves different model families, yielding stronger instruction adherence and much higher visual quality. Beyond the dataset, our analysis highlights that data aligned with a benchmark's distribution can inflate automatic metrics without improving human-judged quality, arguing for preference-oriented curation and evaluation. Our improvements are demonstrated on open models and depend on the base model's capability, and while any static dataset is eventually superseded, we expect the construction pipeline and our analysis to remain useful as stronger generators emerge, since the pipeline can be re-run at a higher quality bar. With Fine-T2I released under an open license, we aim to pave the way for promising modern text-to-image models in the open community.

AI USE STATEMENT

In this work, generative AI is used in two distinct ways. **(i) As part of our methodology.** The construction of Fine-T2I inherently relies on generative and multimodal models, as detailed in Sec. 3: an LLM (LLaMA-3) for controlled prompt synthesis, a fine-tuned prompt enhancer for multi-granularity prompts, strong text-to-image generators (Z-Image and FLUX2) for image synthesis, and reasoning VLMs together with LLaMA-Guard for text–image alignment, attribute, and safety filtering. These uses are integral to the dataset and are documented in the corresponding sections. **(ii) As a writing aid.** The authors additionally used general-purpose AI assistants for minor language polishing and formatting of the manuscript. All research questions, dataset design, experiments, analyses, and claims were conceived and verified by the authors. We have reviewed all AI-assisted text and take full responsibility for the final content of this work, including all text, claims, and released artifacts.

ETHICS STATEMENT

Fine-T2I combines synthetic images with licensed real images. We discuss data licensing and redistribution, the privacy of individuals who may be depicted, content-safety filtering, and our notice-and-takedown policy in the appendix (see the *License and Data Usage* paragraph). Safety filtering with LLaMA-Guard-3 is applied during construction (Sec. 3.1.3) to remove unsafe or non-compliant content, and the dataset is released strictly for non-commercial research use. Beyond the points discussed there, we are not aware of additional ethical concerns.

REPRODUCIBILITY STATEMENT

We provide a complete description of the data-construction pipeline in Sec. 3, covering prompt generation, semantic de-duplication, prompt filtering, image generation, and reasoning-VLM-based text–image verification. The exact system prompt used for final filtering is given in Fig. 11, and detailed distribution statistics, prompt templates, and pre-defined resolutions are reported in Appendix A. The dataset, together with its prompts and annotations, is released under an open license to support full reproducibility of our results.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 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.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 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.

Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yunqi Hong, Kuei-Chun Kao, Hengguang Zhou, and Cho-Jui Hsieh. Understanding reward hacking in text-to-image reinforcement learning. *arXiv preprint arXiv:2601.03468*, 2026.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

jackyhate. text-to-image-2m. `https://huggingface.co/datasets/jackyhate/text-to-image-2M`, 2024.

Kuei-Chun Kao, Daixuan Huo, Yuanhao Ban, and Cho-Jui Hsieh. Autorubric-t2i: Robust rule-based reward model for text-to-image alignment. *arXiv preprint arXiv:2605.17602*, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025a.

Xu Ma, Peize Sun, Haoyu Ma, Hao Tang, Chih-Yao Ma, Jialiang Wang, Kunpeng Li, Xiaoliang Dai, Yujun Shi, Xuan Ju, et al. Token-shuffle: Towards high-resolution image generation with autoregressive models. *arXiv preprint arXiv:2504.17789*, 2025b.

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, 2025c.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

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.

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.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36:49659–49678, 2023.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Guo-Hua Wang, Liangfu Cao, Tianyu Cui, Minghao Fu, Xiaohao Chen, Pengxin Zhan, Jianshan Zhao, Lan Li, Bowen Fu, Jiaqi Liu, et al. Ovis-image technical report. *arXiv preprint arXiv:2511.22982*, 2025a.

Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, et al. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*, 2025b.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. Rewarddance: Reward scaling in visual generation. *arXiv preprint arXiv:2509.08826*, 2025b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

A Detailed Distribution analysis

In this section, we provide a more detailed analysis of our Fine-T2I dataset. Table 5 summarizes style statistics for our synthetic sets. To further characterize the data, Fig. 8 shows the prompt-length distribution. Original prompts are generally short (typically under 50 words), whereas enhanced prompts are substantially longer and span a wider range of lengths. We observe the same trend in both the synthetic sets and the curated real-image set. Finally, we report detailed statistics for categories and tasks in Table 6 and Table 7, respectively. For transparency, we also report our pre-defined resolutions and aspect ratios for creating synthetic images in Table 8.

To ensure diversity in prompt structure and template, we design five distinct instructions (which are added to the system prompts) that guide LLaMA3 during prompt generation. These templates are listed as follows:

- a simple structure, e.g., a photo of xxx, an image of xxx, a picture of xxxx, etc.
- as if the user is making a request, for example, please help generate xxx, Could you please help provide an image xxxx, generate an image of, etc.
- one or multiple sentences.
- combinations of sentence(s) and words (e.g., one or more sentences, + 8K resolution, photorealistic, natural, etc.), etc.
- simple combination of words (e.g., dog, flying, over sea, etc.).

Styles	**PE-AR**	**PE-AS**	**PO-AR**	**PO-AS**	**Total**
3D & CGI	60309	58141	62090	47831	228371
Abstract & Artistic	143509	138761	139430	109619	531319
Anime	189551	183866	189392	143096	705905
Cartoon & Illustration	134538	132375	140222	112841	519976
Cultural & Folk Art	141827	128671	161804	123120	555422
Futuristic & Sci-Fi	92631	92110	86677	59929	331347
Graphic& Digital	210362	200629	221038	175241	807270
Others	36275	34006	37378	29387	137046
Traditional/modern Art	155099	149882	159576	125033	589590
Vintage & Retro	124268	114807	151822	115733	506630
General & Photorealistic	327223	305005	337069	263520	1232817

Table 5: Detailed statistics for styles in our Fine-T2I four synthetic subsets. “PE" indicates enhanced prompt, “PO" indicates original prompt, “AR" means random pre-defined aspect ratios, and “AS" means square aspect ratio. This table shows the detailed sample numbers belonging to each pre-defined style.

B Problems, Ambiguity, Expectation

Diversity of generated Prompt During prompt synthesis, we observed severe redundancy. Prompts generated across different batches and even across different GPUs were often highly similar, and in many cases even identical. Assigning different random seeds for each GPU did not help much. As a mitigation, we increased the LLM sampling temperature to encourage stochasticity. However, this also raised the variance of outputs and occasionally produced malformed generations that were not valid prompts (which we filtered out later). We also found that conditioning on predefined attributes such as style and category further alleviated the issue, but substantial duplication persisted and we did not find a principled way to eliminate it. Consequently, the deduplication stage alone removed nearly 90% of the generated candidates, leading to significant wasted computation.

Attributes alignment with prompt Although we apply attribute-based filtering in Sec. 3.1.3, we do not guarantee the attributes (category, style, and tasks) can be perfectly aligned with the resulting prompt. In practice, some attribute combinations are internally inconsistent or overly constrained. For example, we may request an extremely concise prompt (*e.g.*, less than 10 words)

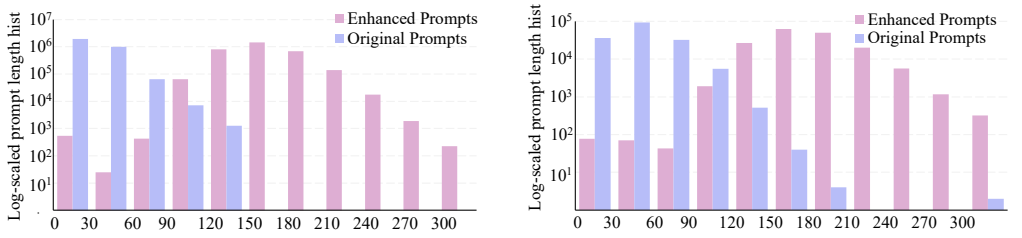
Category		PE-AR	PE-AS	PO-AR	PO-AS	Total
Design	Arts	59722	56909	67021	51893	235545
	Cartoon	21975	22994	25973	20517	91459
	Creative	3734	3593	4245	3267	14839
	Logos	3434	3616	3240	2489	12779
	Posters	28819	27444	28142	21159	105564
	Rare cases	4963	4736	5654	4050	19403
	Slides	19475	18751	17708	13326	69260
	UI/UX	28266	27310	28417	20424	104417
Nature	Animals	59710	57187	66406	51122	234425
	Cityscape	64063	58235	71190	52812	246300
	Creative	4540	4336	4703	3694	17273
	Food	61647	56695	77209	61570	257121
	Indoor	47106	43443	57773	45450	193772
	Landscape	35308	33341	35869	28995	133513
	Objects / Entities	122078	115692	131568	102185	471523
	Plants	34259	32686	36440	29212	132597
	Rare cases	6330	5869	6187	4674	23060
People	Activities	149561	136908	174234	136109	596812
	Age	61683	57572	67566	54927	241748
	Creative	11868	11057	13494	10478	46897
	Emotions	98799	93809	101281	82550	376439
	Portrait	219778	209840	242322	197778	869718
	Rare cases	10186	9289	8822	6290	34587
	Sports	42077	38626	47790	33910	162403
Rare	Rare cases	104062	100557	105787	76979	387385
Text rendering	Creative	5466	5425	4755	3598	19244
	Handwriting	13841	13614	11586	8692	47733
	Long text	74115	70762	43075	29381	217333
	Rare cases	3938	3770	3172	2209	13089
	Short text	111131	110970	101908	77247	401256
	Stylized text	23797	24374	20920	15560	84651
	Text in scene	79861	78843	72041	52803	283548

Table 6: Detailed statistics for categories in our Fine-T2I four synthetic subsets.

Tasks	PE-AR	PE-AS	PO-AR	PO-AS	Total
Colors	84885	81386	89151	70316	325738
Counting	83826	77285	76807	56341	294259
Position	101398	96048	103106	80280	380832
Reasoning	127420	118657	136928	106034	489039
Colors + Counting	29754	27932	25489	18589	101764
Colors + Position	40304	38229	40570	31972	151075
Colors + Reasoning	46128	43543	48277	37638	175586
Counting + Position	37795	35262	33496	24721	131274
Counting + Reasoning	38366	35129	34176	24983	132654
Position + Reasoning	49698	46546	50872	39317	186433
Not Specified	976018	938236	1047626	815159	3777039

Table 7: Detailed statistics for tasks in our Fine-T2I four synthetic subsets. For general generation, over half of the prompts are not assigned a particular task.

while simultaneously requiring a long textual rendering description, or impose strong multi-factor constraints such as UI/UX design + cultural style + counting and colors tasks. Such contradictory or complex conditions can prevent the attributes from being faithfully expressed in the generated prompt. We therefore recommend treating the attributes as soft metadata when using the dataset. They are useful for reference and analysis but may not be accurate, while we ensure the text–image pairs remain reliable for training and evaluation. If stricter attribute adherence is required, one can apply an additional LLM-based reasoning filter (as in our later pipeline stages in Sec. 3.1.6), but this typically results in substantially more aggressive filtering.



(a) synthetic sets prompt length distribution. (b) curated real-image set prompt length distribution.

Figure 8: Prompt length distribution analysis.

	Aspect ratio	Resolution
Square	1:1	[512, 512]
	1:1	[768, 768]
	1:1	[1024, 1024]
	1:1	[1536, 1536]
	1:1	[2048, 2048]
	1:1	[2560, 2560]
Landscape	4:3	[2048, 1536]
	4:3	[1024, 768]
	4:3	[1600, 1200]
	3:2	[1152, 768]
	3:2	[1536, 1024]
	16:9	[1280, 720]
	16:9	[2048, 1152]
	16:9	[2560, 1440]
	5:4	[1280, 1024]
	5:4	[1920, 1536]
Portrait	3:4	[768, 1024]
	3:4	[1200, 1600]
	3:4	[1536, 2048]
	2:3	[768, 1152]
	2:3	[1024, 1536]
	9:16	[720, 1280]
	9:16	[1152, 2048]
	9:16	[1440, 2560]
	4:5	[1024, 1280]
	4:5	[1536, 1920]

Table 8: Pre-defined resolutions and aspect ratios for generating synthetic images.

Ambiguity of aesthetics and human preference What people call a "high-quality" image in text-to-image generation has been changing quickly. A few years ago, highly stylized, visually appealing generations were often considered higher quality. Instead, people now prefer realistic, coherent, and photograph-like generations. This makes "aesthetics" an inherently ambiguous target for dataset collection, and a single automatic aesthetic score (or even a combination of multiple scores) can disagree with what humans actually prefer. Fig. 9 shows a representative example. The top row enjoys a higher aesthetic score, yet people may believe the bottom row is better because it looks more natural. Given this mismatch, we do not treat any aesthetic metric as ground truth. Instead, we follow a pragmatic criterion aligned with current preferences, using the quality signals embodied by strong image generators when generating our synthetic sets, and we trust the taste of professional photographers when collecting our curated real-image set.

Expectation for better image evaluation Building an extremely high-quality fine-tuning dataset required many, sometimes even cumbersome filtering steps. We rely on heavy filtering because there is still no strong, robust, widely accepted indicator for assessing the quality of an image or an image–text pair. We expect that a strong and reliable evaluation metric would substantially simplify this process, and we believe such a metric may benefit from the reasoning capabilities of VLMs.



Figure 9: Aesthetics scores may be ambiguous and may not be aligned with human preference. Indeed, aesthetics always change over time and can be biased. Please zoom in to see details.

C Detailed studies

We first provide our system prompt for the Qwen-VL thinking model in Fig. 11, which checks text–image alignment and image quality against criteria explicitly specified in the system prompt.

We also provide samples from different fine-tuning datasets, as shown in Fig. 12. Instead of cherry-picking specific images, we randomly select a tar file from each dataset and visualize the first several samples. Surprisingly, we observe that some samples from other datasets should not even be included in the fine-tuning set. They are low-resolution, low-quality, and exhibit poor text–image alignment, *etc*., which is far below our expectations for fine-tuning text–image pairs. This also suggests that both the open community and industry need to put more effort into dataset quality, which is often overlooked.

Next, we show fine-tuning examples on different datasets and present the results in Fig. 10. During fine-tuning, we observe that BLIP3o-60k achieves the lowest training loss, but our Fine-T2I clearly delivers the best fine-tuning results. This can be explained by the data characteristics. Images in BLIP3o-60k are generally structurally simple and clear, with limited background and details. This makes it easier for the model to find shortcuts during optimization, but does not indicate better generation performance.

Human Evaluation Protocol We recruit more than 20 anonymous annotators, and each comparison is judged independently by 4–5 of them; unreliable annotators are removed through quality-control checks, and all participation is with informed consent. For the with/without fine-tuning study (Fig. 6), each prompt yields an A/B pair that annotators rate *separately* on two dimensions, text–image alignment and overall visual quality, choosing *A better*, *B better*, or *Tie*; the left/right order is randomized and annotators are blind to which side is fine-tuned, and we ask them to use *Tie* sparingly. For the cross-dataset comparison (Fig. 7), annotators pick the single best image among the three systems (LlamaGen fine-tuned on T2I-2M, BLIP3o-60k, and Fine-T2I), where each system contributes the same number of generations and the display order is randomized. All reported win rates are averaged over prompts and annotators.

License and Data Usage We release Fine-T2I under an open license, strictly for non-commercial research purposes, to support the study of text-to-image generation. The dataset consists of two parts with different provenance. The *synthetic* images are generated by us using publicly released models (Z-Image and FLUX2); we release these images together with all prompts and annotations produced in this work, and their downstream use remains subject to the corresponding upstream model licenses. The *curated real* images are collected from creator-driven platforms (Pexels, Pixabay, and Unsplash-

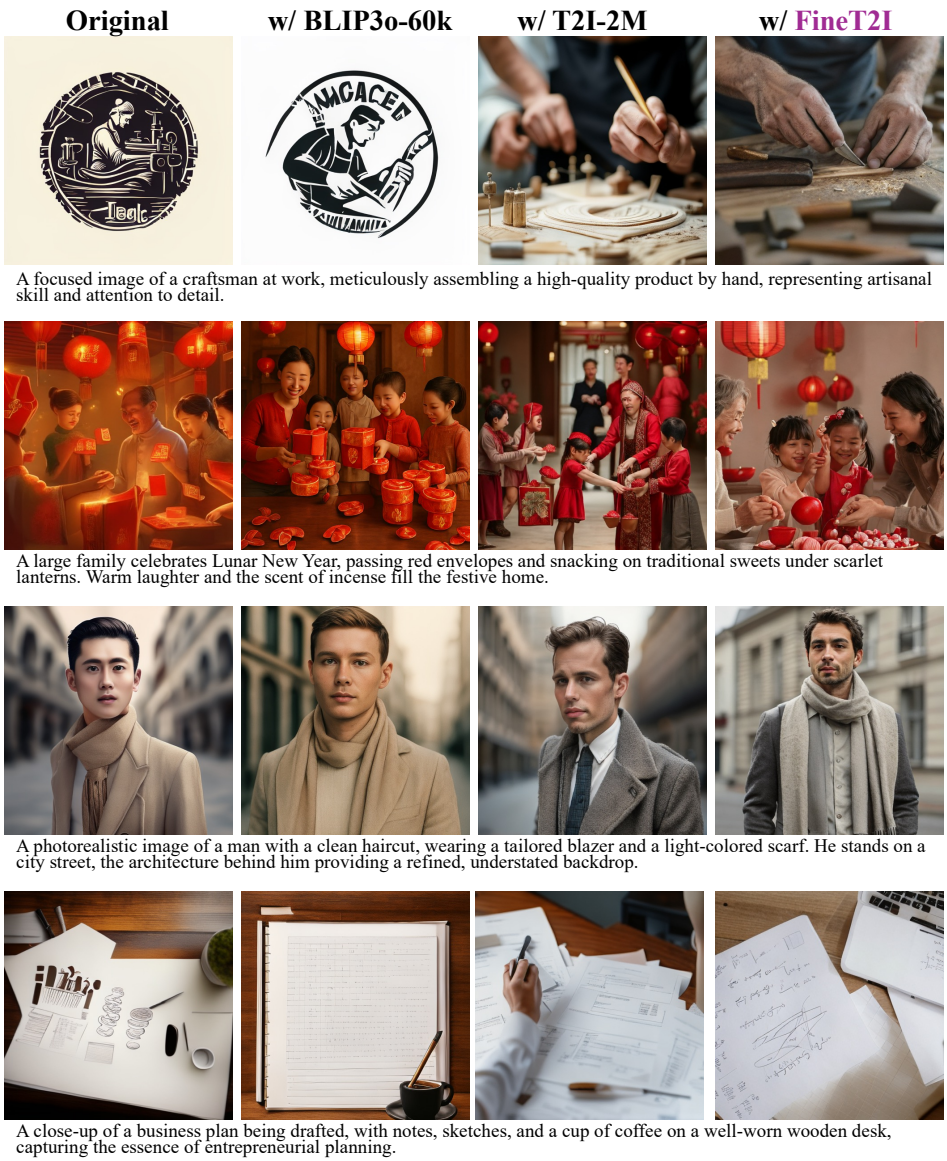


Figure 10: Visual comparison of LlamaGen fine-tuned on different datasets. We randomly pick examples from generated images for visualization.

Lite); for this subset, we include only images for which we obtained the right to use and redistribute for research purposes, either under the applicable content license or through explicit permission from the platform or the rights holders, and we retain per-image source attribution. As some images may depict identifiable individuals, we release no additional personal information and apply content-safety filtering within our pipeline (Sec. 3.1.3). Finally, we maintain a notice-and-takedown process and will promptly remove any image upon a valid request from a rights holder, or whenever we find that it is not properly authorized.

System Prompt

You are a Strict Visual Quality Auditor. Your sole task is to perform a binary pass/fail audit on text-image pairs based on absolute prompt alignment and technical flawlessness.

Evaluation Criteria:

I. Semantic & Text-Image Alignment

- Object Completeness: Every entity, person, or item required in the prompt must be clearly visible and identifiable. Missing any requested element results in an immediate 'False'.
- Quantity Precision: The number of objects in the image must exactly match the count specified in the text. - Text Rendering (OCR): Any text, words, or letters requested in the prompt must be rendered with 100 spelling accuracy, correct font style (if specified), and zero character distortion.
- Attribute & Color Consistency: All specified colors, materials, sizes, and specific properties of objects must be strictly followed.
- Spatial & Relational Logic: Objects must be in the exact positions described . Actions/interactions between objects must be logical.
- Negative Constraints: If the prompt specifies what not to include, the image must not contain those elements.

II. Technical & Aesthetic Quality

- Anatomical Integrity: Zero tolerance for "AI hallucinations" in biological structures. This includes extra/missing fingers, limbs, distorted faces, unnatural joints, or merged body parts, etc.
- Physical & Geometric Logic: Objects must follow the laws of physics (unless the prompt says otherwise). No floating objects, impossible perspectives, unrealistic interactions, or "melting" textures where surfaces bleed into each other.
- Image Artifacts: Zero tolerance for unintended blurring, watermarks, signature-like scribbles, garbled characters, distorted characters or text, etc.
- Style & Medium Fidelity: The image must perfectly embody the requested style (e.g., "CGI," "Oil Painting," "Macro Photography"). If the style is inconsistent across the frame, it is a failure.

Decision Logic:

- Output 'True' ONLY if the image is a perfect realization of the prompt with zero technical defects.
- Output 'False' if there is even one minor discrepancy.
- Extra details are permitted only if they enhance the scene without violating the prompt or the laws of physics.

Constraint:

- Please be concise but thorough in your thinking.
- You must think step-by-step to verify every detail.
- Output ONLY the word 'True' or 'False' as your final answer.

Figure 11: Detailed system prompt used for final text-image pair filtering, for both text-image alignment and aesthetic quality.

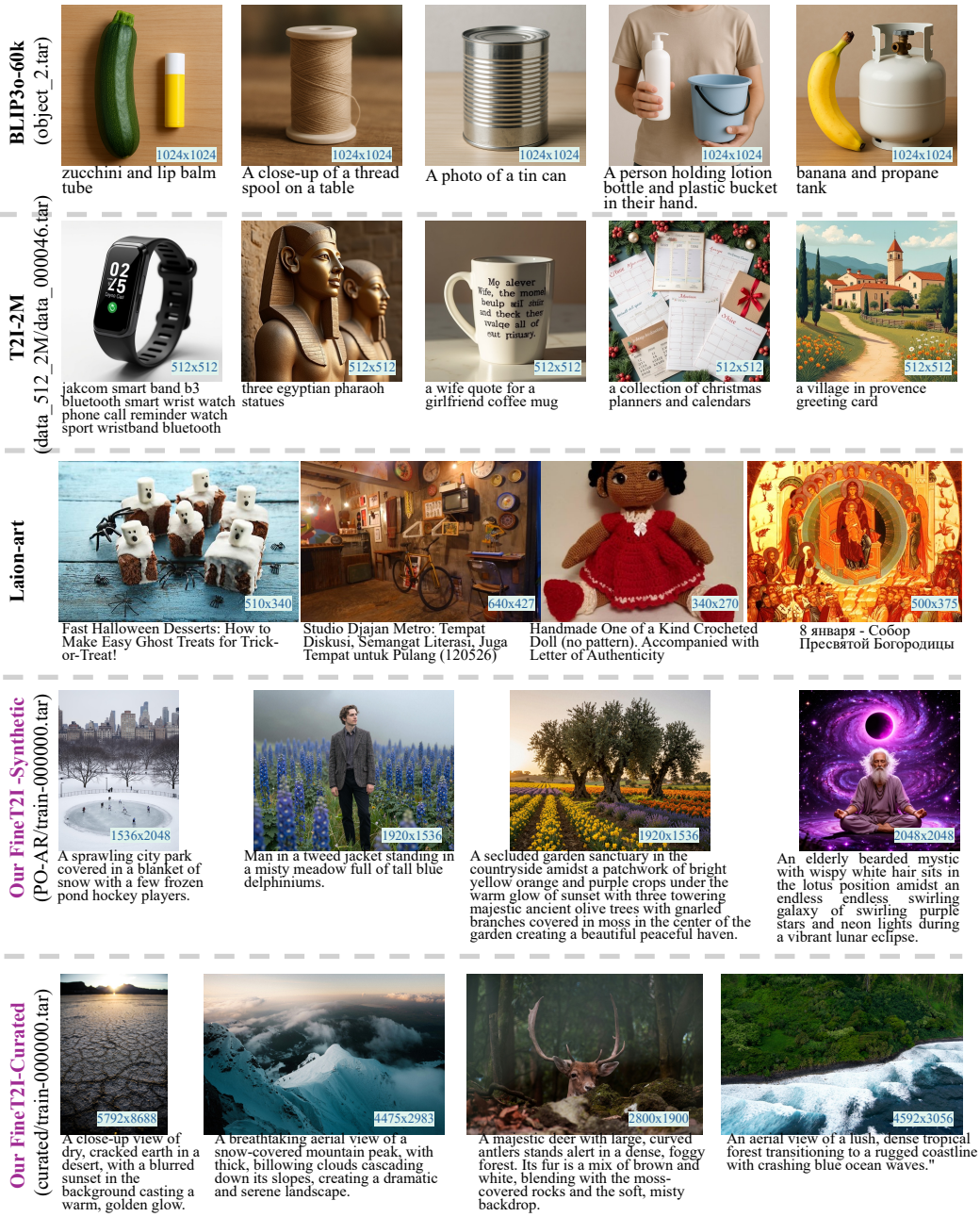


Figure 12: Comparison of text-image pairs from different fine-tuning datasets. We randomly select one tar file from the dataset, and list the first examples as a comparison, without any cherry-picking. The resolution is marked bottom-right. Clearly, our Fine-T2I presents the best alignment, visual quality, and high-resolution, *etc*.