TreeReward: Improve Diffusion Model via Tree-Structured Feedback Learning

Jiacheng Zhang*
Sun Yat-sen University
Guangzhou, Guangdong, China

Jie Wu†
ByteDance
Shenzhen, Guangdong, China

Huafeng Kuang
ByteDance
Shenzhen, Guangdong, China

Haiming Zhang
Chinese University of Hong Kong(Shenzhen)
Shenzhen, Guangdong, China

Yuxi Ren
ByteDance
Shenzhen, Guangdong, China

Weifeng Chen
Sun Yat-sen University
Guangzhou, Guangdong, China

Manlin Zhang
Sun Yat-sen University
Guangzhou, Guangdong, China

Xuefeng Xiao
ByteDance
Beijing, China

Guanbin Li†
Sun Yat-sen University
Shenzhen, Guangdong, China

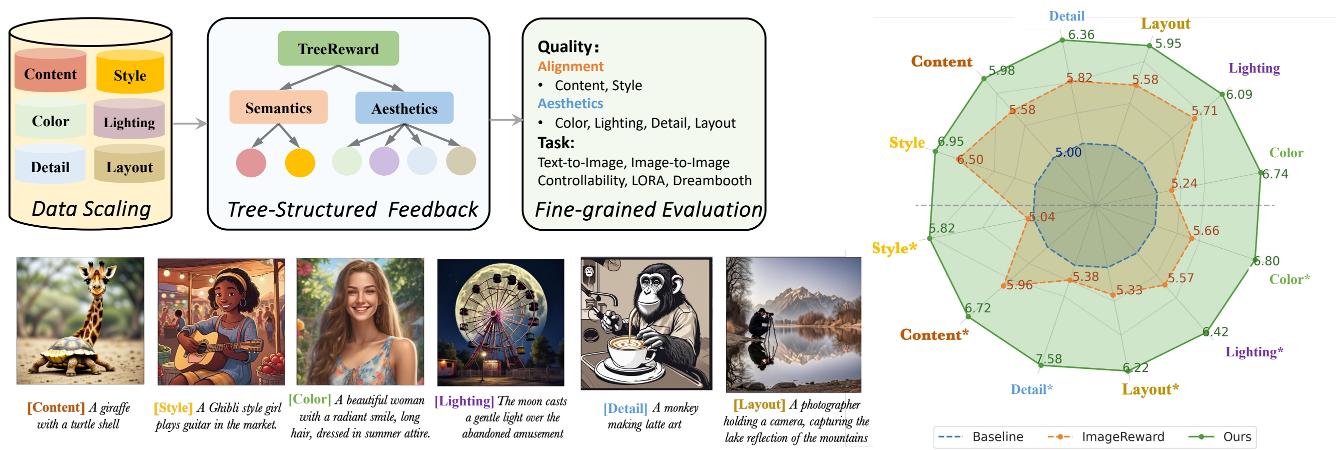

Figure 1: We propose *TreeReward*, including feedback data scaling up, and tree-structured feedback tuning, to achieve fine-grained feedback learning for diffusion models and achieve superior performance under various fine-grained evaluation settings. We consider the fine-grained aspects including Content, Style, Color, Lighting, Detail, and Layout. The radar chart exhibits the fine-grained performance improvement comparison over the baseline (SD1.5). In this chart, the baseline value is anchored at 5, and the value range is rescaled from [-5, 5] to [0, 10] accordingly.

Abstract

Recently, there has been significant progress in leveraging human feedback to enhance diffusion-based image generation, garnering considerable interest and attention. However, existing methods fail to achieve a fine-grained performance boost for the following challenges: i) insufficient amount of fine-grained feedback data; ii) lack of effective fine-grained feedback learning framework; To tackle these challenges, we present **TreeReward** to facilitate the fine-grained feedback optimization for diffusion models. Specifically, to address the limitation of the fine-grained feedback data, we first design a novel "AI + Expert" feedback data construction pipeline, yielding about 2.2M high-quality feedback dataset encompassing six fine-grained dimensions at a relatively low cost. Built upon this dataset, we introduce a tree-structure reward model to exploit the fine-grained feedback data efficiently and provide tailored optimization during feedback learning. We validate the feedback learning performance of our method across different fine-grained dimensions and various downstream tasks. Extensive experiments on both

*Both authors contributed equally to this paper. Work done during internship at ByteDance. Contact Email: zhangjch58@mail2.sysu.edu.cn
†Corresponding author. Email: liguanbin@mail.sysu.edu.cn

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680610

Stable Diffusion v1.5 (SD1.5) and Stable Diffusion XL (SDXL) demonstrate the effectiveness of our method in enhancing the general and fine-grained generation and downstream tasks generalization.

CCS Concepts

• **Computing methodologies → Computer vision**.

Keywords

Diffusion Model, Feedback Learning, Tree-Structure Reward

ACM Reference Format:

Jiacheng Zhang, Jie Wu[†], Huafeng Kuang, Haiming Zhang, Yuxi Ren, Weifeng Chen, Manlin Zhang, Xuefeng Xiao, and Guanbin Li. 2024. TreeReward: Improve Diffusion Model via Tree-Structured Feedback Learning. In *Proceedings of Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680610

1 Introduction

Reinforcement learning from human feedback (RLHF)[1, 24, 26], a technique that incorporates human feedback as the supervision signal and aligns the generation model with the human preference via reinforcement learning methods, has achieved significant advancement in the realm of large language models (LLM) as indicated by notable studies such as LLaMA, GPT-4[1, 24, 26, 39, 46] etc. Recently, some works[43, 44] have emerged to explore applying similar methodologies in the diffusion-based image generation field, aligning the generated images with human preferences. For example, DDPO[2] applies the reinforcement learning algorithm PPO[38] to align the diffusion model with the given reward function by treating the diffusion process as a Markov decision process(MDP). ImageReward[44] develop a direct preference tuning framework for latent diffusion model(LDM), which first trains a reward model on the collected human preference dataset to align with human preference then directly fine-tune the LDM with the reward score guidance within certain denoising timestep range. However, despite these advancements, the diffusion models that integrate these human feedback learning methods still exhibit inferior generation quality in some fine-grained dimensions, for example, generating images with a style not consistent with the prompt or images with unattractive composition. This is because these feedback learning methods lack such fine-grained feedback learning signals to achieve more targeted optimization. A natural question arises, *Could we achieve fine-grained feedback learning for LDM?* After looking closely at this question, we find several challenges exist: (i) *Insufficient fine-grained feedback data*: Current methods generally collect the coarse feedback data(better&worse) without distinguishing which fine-grained dimension has cared. Although ImageReward annotates both the text-to-image alignment and aesthetics for the preference dataset, the aesthetic dimension is still highly coarse and abstract, and the dataset volume is also limited(only 123k). Such limitation on the fine-grained preference feedback data makes it inadequate to capture the diverse range of human preferences. Nonetheless, gathering large-scale fine-grained preference data is labor-intensive and costly. (ii) *No effective fine-grained feedback learning framework*: Most existing methods only involve a single reward model, without the effective practice of modeling multiple fine-grained preferences. Although we can simply train a separate reward model for each dimension, it cannot exploit the relationship between different dimensions and is also unscalable when increasing the fine-grained dimension. To tackle these issues, we propose a fine-grained feedback-learning method, ***TreeReward***, for LDM. For the lack of fine-grained feedback data, we consider the commonly six fine-grained dimensions(i.e. style, content, color, lighting, detail, and layout), and design an "AI + Expert" feedback data generation pipeline that incorporates both the automatic data generation and the human preference annotation according to their distinct properties to enable feedback data scale-up in a low-cost manner. Given such a large-scale curated feedback dataset, we further develop a novel tree-structured reward model. It organizes the fine-grained feedback dimension hierarchically according to their affiliation relationship and employs the random sample ensemble training strategy to effectively integrate multiple fine-grained feedback-scoring abilities into a single reward model. Such designs enjoy several merits, first, the prior knowledge of these fine-grained dimensions is encoded via such tree structure and can facilitate the efficient multi-dimension reward model training. second, during reward tuning, the reward scores from all the leaf nodes are aggregated adaptively, offering the case-tailored feedback signal for more effective optimization. Extensive experiments demonstrate the superiority of our method in enhancing the generation performance of both general quality and fine-grained dimensions. Furthermore, it exhibits exceptional performance in downstream tasks, validating its robustness and generalization.

Our contributions are summarized as follows:

- We develop an efficient feedback data curation pipeline in "AI + Expert" fashion to generate large-scale fine-grained feedback data at low cost, finally yielding about 2.2M feedback dataset which is the largest dataset in the field.
- We design an innovative tree-structure reward model to enable efficient multi-dimensional, and fine-grained feedback modeling, and achieve more effective reward tuning.
- Extensive experiments on both SD1.5 and SDXL validate the effectiveness of our method, demonstrating its superior generation optimization and generalization capability.

2 Related Works

2.1 Text-to-Image Generation

As the representative topic in the Artificial Intelligence Generated Content(AIGC)[3, 19, 24, 25, 29, 32, 40, 41, 47, 51] area, text-to-image(T2I) generation, a task that aims to synthesize the image given a textual description, has become a prominent research field with various applications, attracting significant attention compared with the conventional vision tasks[17, 18, 48–50]. In the past decades, the methodology for T2I generation has gone through from auto-regressive model[5, 7, 30], GAN models[8, 13, 14] to diffusion models(DMs)[9, 23, 35, 52]. Among these, diffusion models[4, 11, 21, 23, 29, 33], generating samples via progressive denoising a Gaussian noise, have recently emerged as the defacto mainstream technique for T2I synthesis due to their impressive generation capabilities as indicated by several notable pioneering studies such as DALLE-2[29], ImageGen[36], and Latent Diffusion Models[28, 33]. However, despite these advancements, diffusion-based T2I models still struggle to generate high-quality images across various fine-grained dimensions, for instance, images with rich and harmonious colors as well as a reasonable layout.

2.2 Learning from Human Feedback

To enhance the generation performance of the large generative models, researchers have proposed human feedback learning[1, 6, 24, 26, 45] to utilize the human preference signal to align model performance with human preference, and have already validated its effectiveness in the works like ChatGPT[25]. It first trains a reward model based on the annotated human preference data, then fine-tune the generative models via the reinforcement learning algorithm like PPO[38]. Inspired by the success in the LLM domain, several works have endeavored to incorporate human feedback into the learning process of diffusion models to better understand human preferences. DDPO [2] adopts a reinforcement learning framework to align diffusion model generation with the supervision provided by an additional reward function. Approaches like HPS[42, 43] employ a separate reward model trained on curated human preference datasets to filter eligible preferred data for fine-tuning stable diffusion. Another approach, Reward Weighting[16], utilizes reward-weighted likelihood as the optimization objective. Recently, ImageReward[44] proposes the ReFL training framework to direct fine-tune stable diffusion via a differentiable reward model. While effective, most of these feedback learning methods rely on a general reward model trained on coarse human preference datasets, limiting their ability to provide fine-grained preference guidance. In this paper, we address these limitations by curating a fine-grained human preference dataset. Building upon this dataset, we propose a tree-structured reward model that offers more effective and flexible reward supervision for diffusion models.

3 Preliminary

3.1 Text-to-Image Diffusion Model

Diffusion-based T2I model formulates image generation as a diffusion and denoising process, it generates high-quality images under the text prompt guidance via gradual denoising from Gaussian noise. During training, a sampled image x is first encoded by a pre-trained VAE encoder to derive its latent representation z. Subsequently, random noise is injected into the latent representation through a forward diffusion process, following a predefined schedule $\{\beta_t\}^T$. This process can be formulated as $z_t = \sqrt{\overline{\alpha}_t} z + \sqrt{1-\overline{\alpha}_t}\epsilon$, where $\epsilon \in \mathcal{N}(0,1)$ is the random noise with identical dimension to z, $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\alpha_t = 1 - \beta_t$ are the predefined noise schedule. To achieve the denoising process, a UNet ϵ_θ is trained to predict the added noise in the forward diffusion process, conditioned on the noised latent and the text prompt c. Formally, the optimization objective of the UNet is:

$$\mathcal{L}(\theta) = \mathbb{E}_{z,\epsilon,c,t}[||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t} z + \sqrt{1-\overline{\alpha}_t}\epsilon, c, t)||_2^2]. \tag{1}$$

During inference, starting from a pure Gaussian noise in the latent, it gradually denoises the noisy latent with the noise predicted via the UNet until getting a clean denoised image after decoding.

3.2 Reward Feedback Learning

Reward feedback learning(ReFL) [44] is a preference fine-tuning framework that aims to improve the diffusion model via human preference feedback. It primarily includes two phases: (1) *Reward Model Training* and (2) *Preference Fine-tuning*. In the Reward Model Training phase, human preference data is collected to train a human preference reward model to capture human preferences. More specifically, considering two collected candidate generations with preference annotation, denoted as x_w (preferred generation) and x_l (unpreferred one). The reward model is encouraged to give a higher reward score for the preferred sample x_w while a lower score for the x_l. Formally, the loss used to train the human preference reward model r_θ can be formulated as follows:

$$\mathcal{L}(\theta)_{rm} = -\mathbb{E}_{(c,x_w,x_l)\sim\mathcal{D}}[log(\sigma(r_\theta(c,x_w) - r_\theta(c,x_l)))], \tag{2}$$

where $\mathcal{D}$ denotes the collected preference feedback data, $\sigma(\cdot)$ represents the sigmoid function, and c corresponds to the text prompt. With this training objective, reward model r_θ is optimized to produce a human preference-aligned score. In the Preference Fine-tuning stage, ReFL begins with an input prompt c, initializing a latent variable x_T at random. The latent variable is first progressively denoised until reaching a randomly selected timestep $t \in [t_1, t_2]$ and a denoised image x_0' is directly predicted from x_t. After that, the pretrained reward model in the previous phase is applied to this denoised image, generating the expected reward score $r_\theta(c, x_0')$. The diffusion model is then fine-tuned to increase the reward score for each sample with the objective:

$$\mathcal{L}(\theta)_{refl} = \mathbb{E}_{c\sim p(c)}\mathbb{E}_{x_0'\sim p(x_0'|c)}[-r(x_0', c)]. \tag{3}$$

4 AI + Expert Feedback Data Scaling

An essential challenge to achieving fine-grained feedback learning for the diffusion model lies in the construction of a high-quality dataset of fine-grained human feedback. Although there are already some available feedback datasets, such as the ImageReward[44] and Pickascore[15], these datasets often suffer from coarse feedback annotation and limited data volume. However, collecting large amounts of fine-grained human feedback data is time-consuming and expensive. To tackle this issue, we design an efficient pipeline to enable large-scale fine-grained feedback data construction at low cost via automatic strategies. The core insight of our design is *not all feedback data on each aspect requires accurate human annotation and we can combine the automatic data generation with manual human annotation to reduce the cost of the data construction.* Through this efficient pipeline, we collect about 2.2M feedback data across six fine-grained dimensions, covering both the semantic and aesthetic aspects. The comparison between our collected feedback data and the existing dataset is summarized in Tab.1.

4.1 Feedback Data on Semantics Alignment

One important desired property of a text-to-image diffusion model is semantic alignment, which requires the generated image to align closely with the text prompt in semantics. Instead of treating the semantic alignment as a whole [44], we propose to decouple the

Name	Annotator	Data Format	Prompt Source	Image Source	# Dim.	# Prompt	# Pairs
HPS [43]	Discord users	Top-1 choice	Discord users	Stable Diffusion	1	25k	25k
ImageReward [44]	Expert	Pairwise	DiffusionDB	DiffusionDB	1	9k	137k
PickScore [15]	Web users	Pairwise	Web users	4 Models	1	38k	584k
HPSv2 [42]	Expert	Pairwise	DiffusionDB*	9 Models + real photo	1	108k	798k
Ours	AI + Expert	Pairwise	DiffusionDB* + Web users	15 Models + real photo	6	400k	2.2M

Table 1: Comparison with other datasets. * indicates that the data has been filtered. Our data covers 6 fine-grained dimensions. Compared with the previous largest dataset, our preferred pair data pair has been expanded by 2.75 ×.

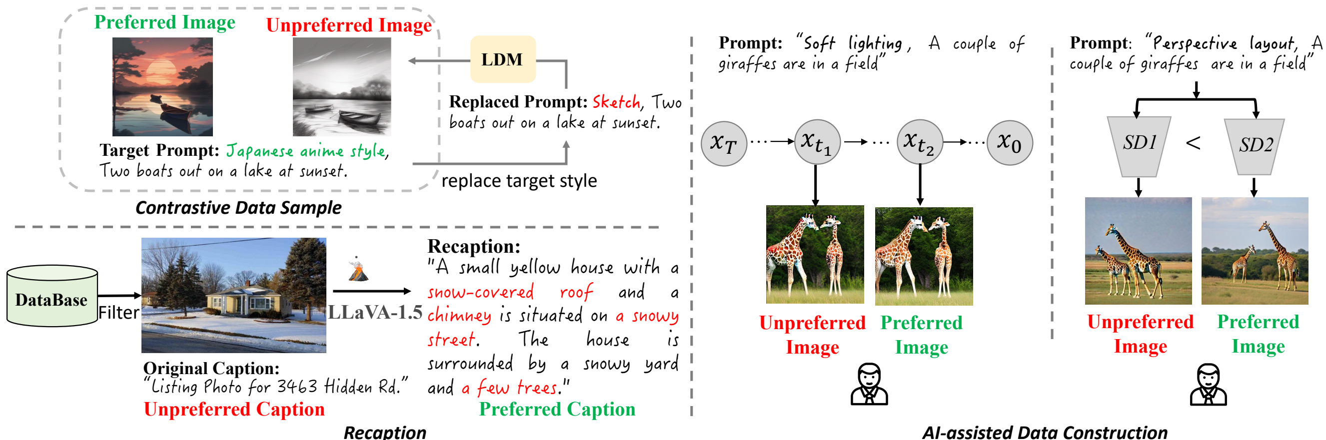

Figure 2: The overview of our "AI + Expert" fine-grained feedback data construction pipeline. We design the automation-assisted data construction strategies for each dimension to enable the fine-grained feedback data scale-up at a low cost.

semantic alignment further into style alignment(consistent stylization) and content alignment(aligned entities and relationship) to aid more targeted alignment in semantics.

Style Alignment Feedback Data. It encourages the model to generate the image in the style specified in the prompt. To achieve this, we develop a *contrastive data sample* strategy to construct the style feedback data as depicted in Fig.2. Specifically, we initially collect a diverse set of approximately 500 commonly used target style words from the real user prompts(e.g. prompt in the JourneyDB[27]). Subsequently, given a prompt containing a particular target style word, we randomly substitute the target style word with another word from the vocabulary. Then, we generated two images with the original prompt and the replaced prompt. Hence, the image generated by the original prompt could serve as a positive sample, whereas the image generated by the randomly substituted style word acts as a negative sample. This is because even though the model cannot fully render the target style, the image generated by the specified style in the prompt will exhibit the stylization toward the target style to a certain degree, which in turn consists of the relative style contrast. To ensure the quality of these feedback data, we exploit the state-of-the-art(SOTA) diffusion model to generate images, such as SDXL[28] and Kindmisky[31]. By employing this methodology, we collect detailed style feedback data that assist in training the diffusion model to accurately capture the desired style.

Content Alignment Feedback Data. Given the input prompt, the diffusion model is expected to generate all the entities and attributes&relationship mentioned in the input prompts accurately. As depicted in Fig.2, we introduce a *recaption* strategy to curate such kinds of feedback data with two steps. i) Identifying Misaligned Examples: We utilized the clip model to identify image-text pairs in the LAION dataset where the clip score fell below a certain threshold. These pairs were considered instances of poor content alignment and sent to re-caption. ii) Generating Detailed Image Descriptions: Given these misaligned text-image pairs, instead of generating a more aligned image(more costly), we inversely re-generate a more aligned prompt via advanced multimodal large language models (MLLMs) such as LLaVA[20]. After that, we considered the image with the original caption and the regenerated caption from the MLLM as the negative and positive sample for content alignment respectively as the regenerated caption tends to be more detailed and accurate with better content alignment.

4.2 Feedback Data on Aesthetics Quality

The aesthetic quality is another critical aspect of the generation performance of the diffusion model. Most of the existing methods only consider coarse aesthetic preference. However, the abstract aesthetic concept contains various dimensions, such as color, lighting, layout, and details, making it challenging for the reward model to grab the aesthetic essence and take the risk of optimization conflict during feedback tuning with such a coarse reward model as analyzed in [41]. To address this limitation, we propose to decouple the aesthetic into more fine-grained dimensions and collect the corresponding feedback data.

Fine-grained Aesthetic Feedback Data. According to the common criterion, we consider several fine-grained dimensions for aesthetics: color, lighting, layout, and details, and collect the corresponding feedback data. Generally, we generate several images with a T2I diffusion model given a particular prompt and then ask the annotator to select the preferred and unpreferred sample. However, the ordinary image pair generated by a text-to-image diffusion model of the same prompt tends to have a similar aesthetic quality, making it challenging to give an assessment and requiring much

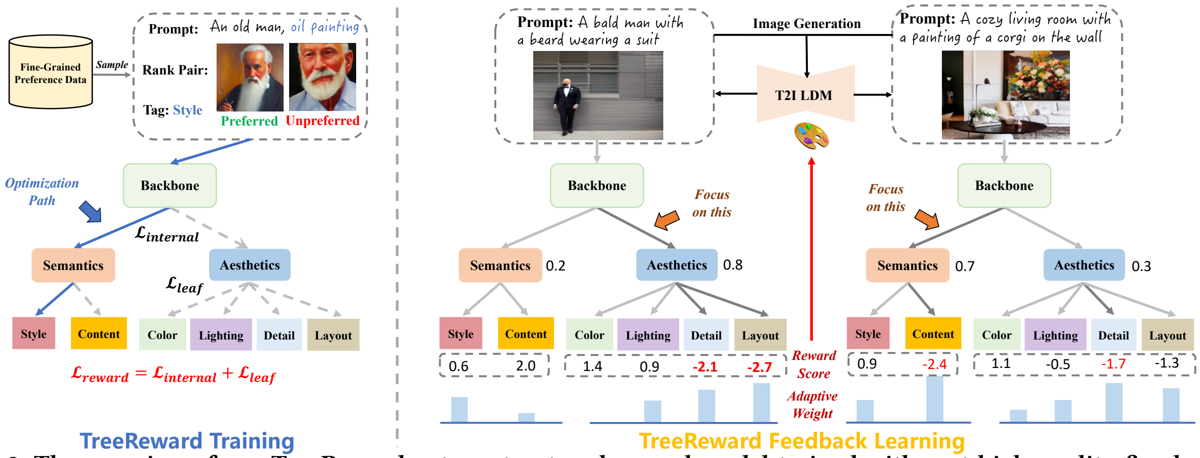

Figure 3: The overview of our TreeReward, a tree-structured reward model trained with vast high-quality fined-grained preference data to facilitate more effective feedback learning for text-to-image generative models.

more time to distinguish the preferred sample. To this end, we propose two strategies to ease the human annotation burden during feedback data generation. (i) To make the generation quality on a particular dimension more prominent, we manually curate a set of trigger words for each aesthetic dimension. For instance, we set the trigger words like "Soft lighting," "Side lighting," and "rim lighting," for lighting dimension. (ii) To reduce the annotation time cost, we manually create the sample with aesthetic differences as depicted in Fig.2. On one hand, we utilize diffusion models with varying generative capabilities to generate the image pair. For example, we take the image generated by SDXL[28] and its improved version Kindminsky[31]. On the other hand, we take the images generated at different denoised timesteps to assist the fast preference annotation. These designs make the candidate images exhibit more clearer performance gap to be captured, facilitating fast annotation. With these two strategies applied, we can achieve efficient fine-grained aesthetic feedback data annotation.

5 Tree-Structured Feedback Learning

Through the proposed "AI + Expert" feedback data construction pipeline, we finally curate a large-scale(~2.2M) fine-grained feedback data spanning six dimensions. The next question is how to utilize these datasets efficiently. There are two simple ways to utilize these datasets: (1) Treat these datasets as a whole and train a global reward model. (2) Train a separate reward model for each dimension. However, these ways either overlook the difference between these dimensions or ignore the inherent prior knowledge about these dimensions, which results in sub-optimal feedback tuning results as shown by our experiments. To this end, we introduce TreeReward, to combine these fine-grained feedback data efficiently and provide fine-grained and adaptive preference guidance during fine-tuning. Fig.3 shows the TreeReward training and preference fine-tuning process, which will be explained in the following sections.

5.1 TreeReward Training

Tree Structure Reward: We take a hard-coded tree structure to organize these preference rewards. Generally, we group these dimensions accordingly as semantic aspect(i.e. style, content) and aesthetic aspect(i.e. color, lighting, detail, layout), and design a hierarchical structure comprised of two internal nodes and six leaf nodes. The role of the internal nodes is to determine which general aspect (semantic or aesthetic) to reward, while the leaf nodes are responsible for providing precise reward scores along fine-grained dimensions. In line with ImageReward[44], we implement TreeReward with the BLIP model as the backbone, which employs ViT-L14 as the image encoder and a 12-layer transformer as the text encoder. For the internal node and leaf node of our TreeReward, we implement them with simple 2× and 4× MLPs, respectively.

Reward Training: We take a *random sample ensemble* strategy to train our TreeReward. Specifically, we first randomly sampled a data point(pair-wise form) from the collected feedback dataset. Then, a reward path is chosen from the root of TreeReward to the leaf node according to the annotated dimension. The target is to train the model to reward the sample correctly along this path, which consists of the internal and leaf node training objectives. Formally, for the internal nodes,

$$\mathcal{L}_{internal}(\theta) = \mathbf{CE}(g, G(x)). \tag{4}$$

Here, $G(\cdot)$ represents the predicted logits of the internal node, and g is the target internal reward label determined by the selected data point, **CE** represents the cross-entropy loss. This objective aims to let the model know which aspect to reward when given a text-image pair. For the leaf node, we expect the leaf node to output the correct reward score. We take a similar way with E.q.2 to optimize the leaf node along a particular reward dimension. Specifically, for the j-th reward local leaf node denoted as R_θ^j, the preference feedback reward data pair is represented as (x_w, x_l). The loss function for the leaf node can be formulated as:

$$\mathcal{L}_{leaf}(\theta) = -\mathbb{E}(x_w, x_l) \sim D_j[\log(\sigma(r_\theta^i(x_w)) - \sigma(r_\theta^i(x_l)))]. \tag{5}$$

where $r_\theta^j(x.)$ is the scalar reward predicted by the j-th leaf node, D^j represents the corresponding fine-grained reward dataset, σ is the sigmoid function. The complete loss for training TreeReward is defined as:

$$\mathcal{L}(\theta)_{RM} = \mathcal{L}_{internal}(\theta) + \mathcal{L}_{leaf}(\theta). \tag{6}$$

Table 2: The quantitative results between the SOTA model and ours in clip and aesthetic score. Among them, * means that the model fine-tuned with JourneyDB.

Model	CLIP Score	Aesthetic Score
SD1.5	25.60	5.41
SD1.5+ImageReward	26.20	5.55
SD1.5+TreeReward	**26.70**	**5.62**
SD1.5*	26.60	5.89
SD1.5*+ImageReward	26.90	5.90
SD1.5*+TreeReward	**27.40**	**5.96**
SDXL	27.28	5.69
SDXL+ImageReward	27.35	5.66
SDXL+TreeReward	**27.39**	**5.84**

It is worth noting that we only optimize one particular path in the TreeReward according to the source of the feedback data for each time, and leave the parameters of other nodes unchanged. However, with the random training data sampling, the parameters of the whole tree will be fully optimized.

5.2 TreeReward Feedback Learning

We adopt the direct preference fine-tuning[44] to exploit the reward score derived from our TreeReward to fine-tune the diffusion model adaptively. Specifically, given prompt and generated image pair (y_i, x_i), we start from the root node, and first calculate the internal node logits and obtain the reward weights w_b for the global reward aspect. Formally, we have:

$$w_b = \frac{e^{G_b(x_i,y_i)}}{\sum e^{G_b(x_i,y_i)}}, w_k = \frac{e^{-r_k(x_i,y_i)}}{\sum e^{-r_k(x_i,y_i)}}, \tag{7}$$

where $G_\theta^b(x)$ with $b \in \{\text{semantic, aesthetic}\}$ represents the prediction output of the internal node, and w_b is the reward weight of along these two aspects. Next, we obtain all the fine-grained reward scores on the leaf node and compute the adaptive weight of each leaf node w_k under the internal node, where k is the leaf node under an internal node. The final reward R is obtained by combining the rewards hierarchically from the root to the leaf:

$$R_{tr}(y_i, g_\theta(y_i)) = \sum_{b=0}^{M} w_b \sum_{k=0}^{N_b} w_k \cdot r_k(x_i, y_i), \tag{8}$$

where N_b is the number of reward leaf nodes under the internal node b, and M is the number of internal nodes. By combining the fine-grained rewards from all nodes, our model can adaptively focus on the reward dimensions that have not been well optimized yet, providing case-tailored preference feedback for the diffusion model via:

$$\mathcal{L}_{reward} = \mathbb{E}y_i \sim y\left[-R_{tr}(x_i, y_i)\right]. \tag{9}$$

Following [44], we also incorporate the naive diffusion pretrain loss as a regularization term:

$$\mathcal{L}_{pretrain} = \mathbb{E}_{(y_i,x_i)\sim D}\left(\mathbb{E}_{\mathcal{E}(x_i),y_i,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_i, t, \tau_\theta(y_i))\|_2^2\right]\right). \tag{10}$$

Therefore, the final training objective is:

$$\mathcal{L} = \mathcal{L}_{pretrain} + \lambda\mathcal{L}_{reward}, \tag{11}$$

where λ is the loss weight and is set to 0.05 by default.

6 Experiments

6.1 Implementation Details

Dataset and Training Setting. We utilize the collected fine-grained preference data to train our TreeReward. We randomly sample a prompt set(∼10w) from DiffusionDB following [44] for preference fine-tuning. We conduct experiments with Stable Diffusion v1.5 and Stable Diffusion XL base 1.0. Additionally, to validate the effectiveness and generalization of our method, we further utilize the JourneyDB [27], a large-scale generated image dataset collected from Midjourney, to fine-tune the base SD1.5 to acquire an improved base diffusion model as our base model.

Evaluation Metrics. We employ the CLIP score[10](ViT-L 14) and Aesthetic score(LAION[37] aesthetic predictor) to assess the general performance of our method on prompt-following and the aesthetic quality, respectively. In addition, we further evaluate the fine-grained optimization performance via a comprehensive user study. Specifically, we first curate 100 prompts for each dimension, which are manually checked to ensure they describe the picture most relevant to the corresponding fine-grained dimension. Subsequently, we ask 10 raters to score what degree the image generated with these prompts by the optimized model is better than the one generated by the original base model. The score spans from -5 to 5, where 5 means the best and -5 represents the worst.

Table 3: 'IR': ImageReward. 'TR': TreeReward.

Setting	Data Volume	Structure	Reward Nums
ImageReward	137K	IR	single
Data-Scale	2.2M	IR	single
Decouple	2.2M	IR	multiple
TreeReward (Ours)	2.2M	TR	single

6.2 Comparision with State-of-the-art

Qualitative Results. We compared our method with ImageReward, the current state-of-the-art SD preference modeling method. It clearly shows that our method exhibits superior performance for learning performance in both semantic alignment and aesthetic quality enhancement. As shown in Fig.4, our method exhibits **superior overall visual quality**. Take SD1.5 and the prompt of "*A dolphin leaps through the waves, set against a backdrop of bright blues and teal hues*" as an example, there are no dolphins in the image generated by SD1.5. The dolphin generated by ImageReward is too small and the waves are blurry. By contrast, both the waves and dolphins generated by TreeReward are rich in detail and highly realistic. As depicted in Fig.6, TreeReward also shows superiority in generating images with **better visual quality in various fine-grained aspects**. For example, only TreeReward generates the correct result for prompt "*a horse without a rider*", while both the base model and ImageReward generate the mismatched content (The riders). And for prompt "*A mountain retreat's spa, zen-inspired, ... overlook forest views*", both SD1.5 and ImageReward present unreasonable layouts for the tables and swimming pool (Too small tables and truncated swimming pool), while TreeReward displays the more aesthetic layout. Note that the ImageReward does not exhibit much improvement when applied to the improved stable diffusion base model which is fine-tuned with JourneyDB. As a comparison, Our TreeReward still delivers notable improvement, which demonstrates the superiority of our method.

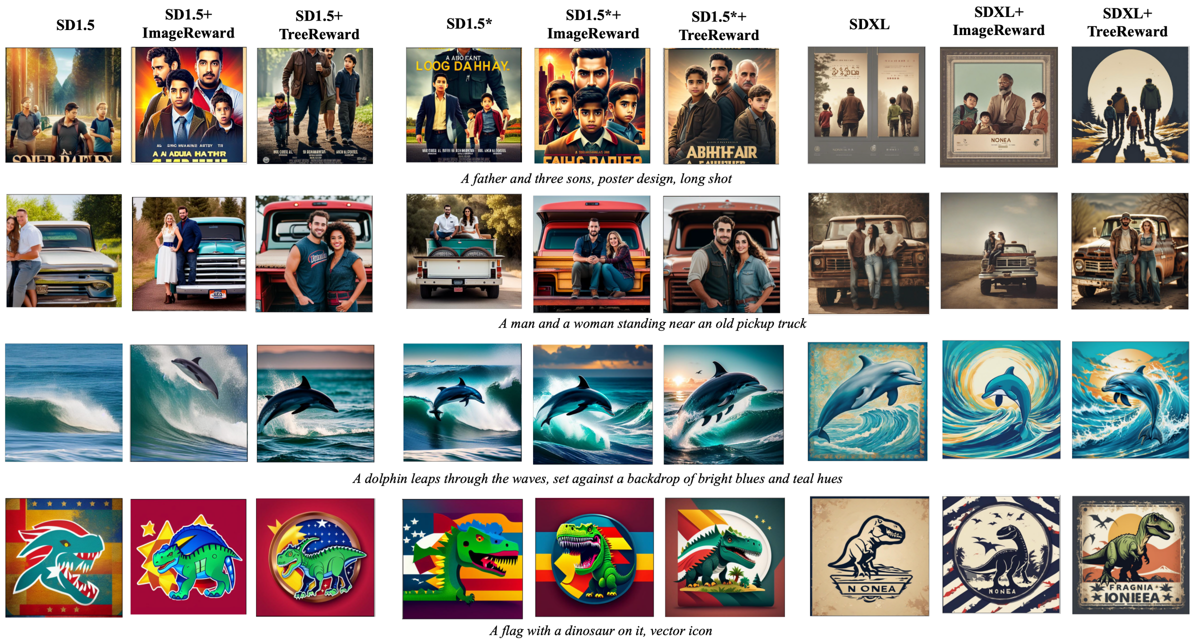

Figure 4: Visual comparison of SOTA models. TreeReward has achieved more excellent results than other competitive methods.

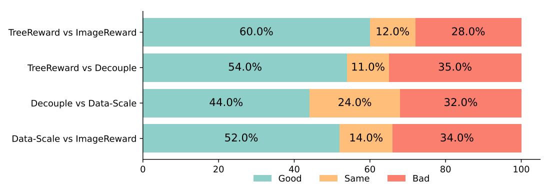

Figure 5: The user study to validate feedback data scaling up and tree-structured reward design. The result is evaluated by 10 raters on the generation of 100 general prompts.

Quantitative results. To evaluate the performance of our method quantitatively, we conducted comparisons using CLIP scores and aesthetic scores, which provide metrics for semantic alignment and aesthetic quality, respectively. The results are presented in Tab. 2. It demonstrates that our method outperforms the baseline model in both semantic alignment and aesthetic quality and also surpasses the performance of ImageReward. For instance, on SD1.5, our method achieved a 2% improvement in semantic alignment compared to ImageReward, along with a 1.2% enhancement in aesthetic quality. These results indicate the superiority of our method in generating images that are not only visually appealing but also semantically aligned. It is worth emphasizing that the score achieved by the SD1.5 model with fine-tuning using journey-db data is higher than that of the SDXL model. This observation underscores the significance of utilizing high-quality fine-tuned data to enhance model performance. Moreover, the fine-grained evaluation is presented in the radar char in Fig.1. It is evident that our TreeReward feedback learning approach significantly outperforms the ImageReward across all the dimensions. Remarkably, our TreeReward approach demonstrates a notable enhancement of 1.5 points in the 'Color' dimension when compared to the ImageReward method on SD1.5. Moreover, when applied on JourneyDB-tuned SD1.5, our TreeReward approach still showcases a significant improvement of 2.2 points in the 'Detail' dimension, surpassing the ImageReward model by a substantial margin.

6.3 Ablation Study

We have conducted a series of ablation experiments to showcase the key contributions of our method, specifically the fine-grained feedback data scaling and the design and tree-structured reward fine-tuning. These experiments encompass several settings: (i) **"Data-Scale"**: We employ the same reward model as ImageReward but utilize our collected feedback data for training the reward model without distinguishing the different fine-grained dimensions. (ii) **"Decouple"**: Instead of training a single reward model, we train separate reward models for each fine-grained dimension using our feedback data and utilize these models for preference fine-tuning simultaneously. (iii) **"ImageReward"**: Preference fine-tuning using the reward model provided by ImageReward. (iv) **"TreeReward"**: Preference fine-tuning using the reward model provided by our TreeReward approach. The detailed comparison between these settings is presented in Tab.3. As illustrated in Fig.5, incorporating more feedback data significantly enhances the performance of preference fine-tuning, resulting in an impressive increase of **18%** compared to the naive ImageReward approach. This finding underscores the importance of gathering a larger quantity of high-quality feedback data, even without considering fine-grained distinctiveness. Building upon this, the decoupling of the reward model for different fine-grained dimensions leads to a further improvement (44% vs 32%). This demonstrates the necessity of the decoupled reward model design, which effectively eliminates potential conflicts in preference tuning as analyzed in [41]. However, training multiple reward models not only results in memory inefficiency but also achieves sub-optimal multiple reward fine-tuning. In comparison, our TreeReward approach leverages fine-grained feedback data hierarchically and rewards in an adaptive manner, offering

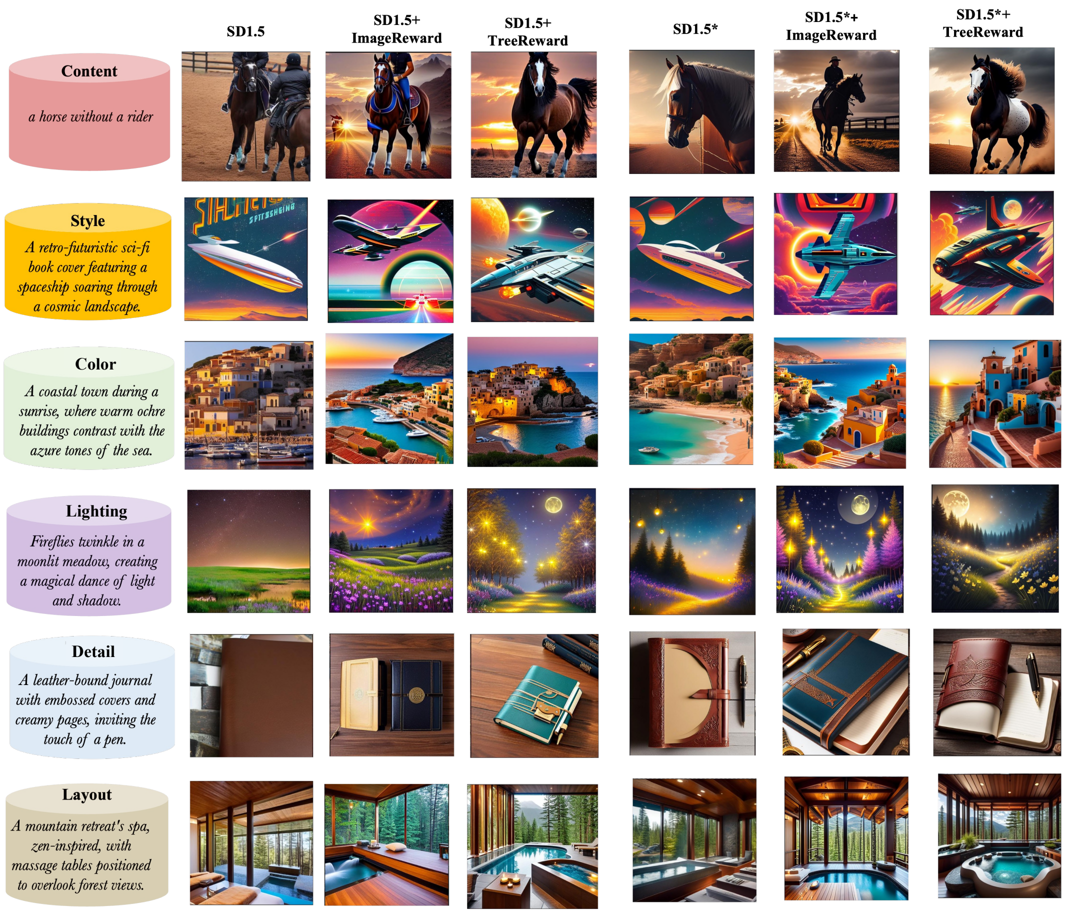

Figure 6: Visual comparison of each fine-grained dimension. * means that the model fine-tuned with JourneyDB.

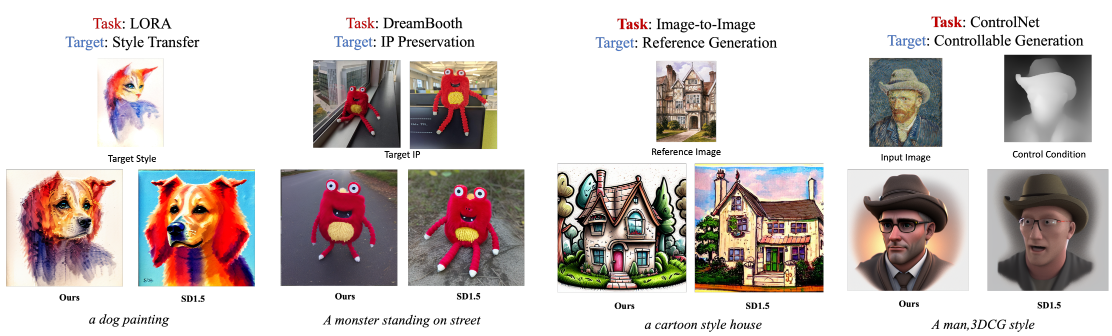

Figure 7: Comparison of visual results on the downstream task with SD1.5 and SD1.5 optimized by our method.

greater flexibility and delivering superior performance compared to naive fine-tuning with multiple reward models (54% vs 35%). By incorporating these two improvements, our method ultimately achieves a **32%** increase in user preference compared to ImageReward, highlighting the significant advantages of our approach.

6.4 Generalization Study

We conduct an extensive study to evaluate the generalization potential of our method in adapting to various downstream tasks, such as LORA [12], DreamBooth [34], Image-to-Image[22], and ControlNet [53]. As illustrated in Fig.7, our model showcases remarkable compacity in style learning, IP preservation, reference generation, and controllable generation.

7 Conclusion

We propose TreeReward, an effective method to achieve fine-grained feedback learning for the diffusion model. It includes an efficient "AI + Expert" fine-grained feedback data construction pipeline, and a tree-structured reward model to impose fine-grained, multi-dimensional, and adaptive reward feedback tuning. Extensive experiments on both SD1.5 and SDXL models demonstrate the superiority of our method in both boosting the general quality and fine-grained generation and at the same exhibiting excellent generalization.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NO. 62322608), in part by the Fundamental Research Funds for the Central Universities under Grant 22lgqb25, and in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2023A01).

References

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL]

[2] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2023. Training Diffusion Models with Reinforcement Learning. arXiv:2305.13301 [cs.LG]

[3] Weifeng Chen, Jiacheng Zhang, Jie Wu, Hefeng Wu, Xuefeng Xiao, and Liang Lin. 2024. ID-Aligner: Enhancing Identity-Preserving Text-to-Image Generation with Reward Feedback Learning. *arXiv preprint arXiv:2404.15449* (2024).

[4] Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs.LG]

[5] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* 34 (2021), 19822–19835.

[6] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767* (2023).

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[9] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10696–10706.

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).

[11] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598 [cs.LG]

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL]

[13] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. 2018. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 266–273.

[14] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[15] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. arXiv:2305.01569 [cs.CV]

[16] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning Text-to-Image Models using Human Feedback. arXiv:2302.12192 [cs.LG]

[17] Jiaming Li, Xiangru Lin, Wei Zhang, Xiao Tan, Yingying Li, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. 2023. Gradient-based sampling for class imbalanced semi-supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16390–16400.

[18] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. 2024. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16678–16687.

[19] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. 2024. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. *arXiv preprint arXiv:2404.07987* (2024).

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 [cs.CV]

[21] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. 2022. More Control for Free! Image Synthesis with Semantic Diffusion Guidance. arXiv:2112.05744 [cs.CV]

[22] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. 2023. Subject-Diffusion:Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. arXiv:2307.11410 [cs.CV]

[23] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741 [cs.CV]

[24] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[25] OpenAI. 2023. Introducing chatgpt. arXiv:2303.08774 [cs.CL]

[26] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

[27] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2023. JourneyDB: A Benchmark for Generative Image Understanding. arXiv:2307.00716 [cs.CV]

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

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[31] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. 2023. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502* (2023).

[32] Yuxi Ren, Jie Wu, Yanzuo Lu, Huafeng Kuang, Xin Xia, Xionghui Wang, Qianqian Wang, Yixing Zhu, Pan Xie, Shiyin Wang, et al. 2024. ByteEdit: Boost, Comply and Accelerate Generative Image Editing. *arXiv preprint arXiv:2404.04860* (2024).

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. arXiv:2208.12242 [cs.CV]

[35] 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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[36] 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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 [cs.CV] https://arxiv.org/abs/2210.08402

[38] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro,

Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[42] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. arXiv:2306.09341 [cs.CV]

[43] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human Preference Score: Better Aligning Text-to-Image Models with Human Preference. arXiv:2303.14420 [cs.CV]

[44] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. arXiv:2304.05977 [cs.CV]

[45] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302* (2023).

[46] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).

[47] Haiming Zhang, Zhihao Yuan, Chaoda Zheng, Xu Yan, Baoyuan Wang, Guanbin Li, Song Wu, Shuguang Cui, and Zhen Li. 2023. GSmoothFace: Generalized Smooth Talking Face Generation via Fine Grained 3D Face Guidance. *arXiv preprint arXiv:2312.07385* (2023).

[48] Jiacheng Zhang, Jiaming Li, Xiangru Lin, Wei Zhang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. 2024. Decoupled Pseudo-labeling for Semi-Supervised Monocular 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16923–16932.

[49] Jiacheng Zhang, Xiangru Lin, Minyue Jiang, Yue Yu, Chenting Gong, Wei Zhang, Xiao Tan, Yingying Li, Errui Ding, and Guanbin Li. 2022. A multi-granularity retrieval system for natural language-based vehicle retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3216–3225.

[50] Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. 2023. Semi-detr: Semi-supervised object detection with detection transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 23809–23818.

[51] Jiacheng Zhang, Jie Wu, Yuxi Ren, Xin Xia, Huafeng Kuang, Pan Xie, Jiashi Li, Xuefeng Xiao, Weilin Huang, Min Zheng, et al. 2024. UniFL: Improve Stable Diffusion via Unified Feedback Learning. *arXiv preprint arXiv:2404.05595* (2024).

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV]