AGFSync: Leveraging AI-Generated Feedback for Preference Optimization in Text-to-Image Generation

Jingkun An[1*], Yinghao Zhu[1*], Zongjian Li[2*], Enshen Zhou[1], Haoran Feng[3], Xijie Huang[1], Bohua Chen[4], Yemin Shi[2], Chengwei Pan[1, 5†]

[1]Beihang University, Beijing, China
[2]Peking University, Beijing, China
[3]Tsinghua University, Beijing, China
[4]Huazhong University of Science and Technology, Wuhan, Hubei, China
[5]Zhongguancun Laboratory, Beijing, China
anjingkun02@gmail.com, pancw@buaa.edu.cn

Abstract

Text-to-Image (T2I) diffusion models have achieved remarkable success in image generation. Despite their progress, challenges remain in both prompt-following ability, image quality and lack of high-quality datasets, which are essential for refining these models. As acquiring labeled data is costly, we introduce AGFSync, a framework that enhances T2I diffusion models through Direct Preference Optimization (DPO) in a fully AI-driven approach. AGFSync utilizes Vision-Language Models (VLM) to assess image quality across style, coherence, and aesthetics, generating feedback data within an AI-driven loop. By applying AGFSync to leading T2I models such as SD v1.4, v1.5, and SDXL-base, our extensive experiments on the TIFA dataset demonstrate notable improvements in VQA scores, aesthetic evaluations, and performance on the HPS v2 benchmark, consistently outperforming the base models. AGFSync's method of refining T2I diffusion models paves the way for scalable alignment techniques.

Project — https://anjingkun.github.io/AGFSync

1 Introduction

The advent of Text-to-Image (T2I) generation technology represents a significant advancement in generative AI. Recent breakthroughs have predominantly utilized diffusion models to generate images from textual prompts (Rombach et al. 2022a; Betker et al. 2023; Podell et al. 2023; Zhang, Rao, and Agrawala 2023; Zhou et al. 2024b). However, achieving high fidelity and aesthetics in generated images poses challenges, including deviations from prompts and inadequate image quality (Zhang et al. 2023). Addressing these challenges requires enhancing diffusion models' ability to accurately interpret detailed prompts (prompt-following ability (Betker et al. 2023)) and improve the generative quality across style, coherence, and aesthetics.

Efforts to overcome these challenges span dataset, model, and training levels. High-quality text-image pair datasets, as proposed in the data-centric AI philosophy, can significantly improve performance (Zhou et al. 2024a). Therefore a high-quality image caption and its corresponding image pair dataset is crucial in training (Betker et al. 2023).

At the model architecture level, advancements include the optimization of cross-attention mechanisms to improve model compliance (Feng et al. 2023). These efforts, both at the dataset and model architecture levels, follow the traditional training paradigm of using elaborately designed models with specific datasets. In contrast, in the training domain, strategies inspired by the success of large language models, such as OpenAI's ChatGPT (OpenAI 2023), include supervised finetuning (SFT) and alignment stages. With a pretrained T2I diffusion model, enhancing the model for better image quality can be approached in either the SFT stage or the alignment stage. The former approach, as seen in the latest work DreamSync (Sun et al. 2023), finetunes the diffusion model through a selected image selection procedure where a Vision-Language Model (VLM) (Achiam et al. 2023; Qin et al. 2023; Zhou et al. 2024c; Qin et al. 2024) evaluates and then selects high-quality text-image pairings for further finetuning. However, DreamSync exhibits a lower prompt generation conversion rate and is limited by the intrinsic capabilities of the diffusion model, leading to uncontrollable data distribution in the finetuning dataset. The latter approach, DPOK (Fan et al. 2024), DDPO (Black et al. 2023), and DPO (Rafailov et al. 2023) use reinforcement learning for alignment, while Diffusion-DPO (Wallace et al. 2023) applies Direct Preference Optimization (DPO) for model alignment, modifying the original DPO algorithm to directly optimize diffusion models based on preference data. Yet, it only focuses on evaluating image quality from one aspect. Furthermore, existing methods mostly depend on extensive, quality-controlled labeled data.

Addressing these requires a cost-effective, low-labor approach that minimizes the need for human-labeled data while considering multiple quality aspects of images. Leveraging AI in generating datasets and evaluating image quality can fill these gaps without human intervention. Through generating diverse textual prompts, assessing generated images, and constructing a comprehensive preference dataset,

*These authors contributed equally.
†Corresponding author.

AGFSync epitomizes the full spectrum of AI-driven innovation—ushering in an era of enhanced data utility, accessibility, scalability, and process automation while simultaneously mitigating the costs and limitations associated with manual data labeling.

More specifically, AGFSync aligns text-to-image diffusion models via DPO, with multi-aspect AI feedback generated data. The process begins with the preference candidate set generation, where LLM generates descriptions of diverse styles and categories, serving as high-quality textual prompts. Candidate images are then generated using these AI-generated prompts, therefore constructing candidate prompt-image pairs. Image evaluation and VQA data construction follow, using LLM to generate questions related to the composition elements, style, etc., based on its initial prompts. VQA scoring is conducted by inputting these questions into the VQA model to assess whether the diffusion model-generated images aesthetically follow the prompts, calculating accuracy as the VQA score. With combined weighted scores of VQA, CLIP, and aesthetics filtering, the preference pair dataset is established within the best and worst images. Finally, DPO alignment is applied to the diffusion model using the constructed preference pair dataset. The entire process leverages the robust capabilities of VLMs without any human engagement, ensuring a human-free, cost-effective workflow.

Our contributions are summarized as follows: **(1)** We introduce a dataset composed of 45.8K AI-generated prompt samples and corresponding SDXL-generated images, each accompanied by question-answer pairs that validate the image generation's fidelity to textual prompts. This dataset not only propels forward the research in T2I generation but also embodies the shift towards higher data utilization, scalability, and generalization, signifying a breakthrough in mitigating the unsustainable practices of manual data annotation. **(2)** Our proposed framework AGFSync, aided by multiple evaluation scores, leveraging DPO finetuning approach, introduces a fully automated, AI-driven approach, which elevates fidelity and aesthetic quality across varied scenarios without human annotations. **(3)** Extensive experiments demonstrate that AGFSync significantly and consistently improves upon existing diffusion models in terms of adherence to text prompts and overall image quality, establishing the efficacy and transformative potential of our AI-driven data generation, evaluation and finetuning framework.

2 Related Work

2.1 Aligning Diffusion Models Methods

The primary focus of related work in this area is to enhance the fidelity of images generated by diffusion models in response to text prompts, ensuring they align more closely with human preferences. This endeavor spans across dataset curation, model architecture enhancements, and specialized training methodologies.

Dataset-Level Approaches: A pivotal aspect of improving image generation models involves curating and finetuning datasets that are deemed visually appealing. Works by (Podell et al. 2023; Rombach et al. 2022a) utilize datasets rated highly by aesthetics classifiers (Schuhmann 2022) to bias models towards generating visually appealing images. Similarly, Emu (Dai et al. 2023) enhances both visual appeal and text alignment through finetuning on a curated dataset of high-quality photographs with detailed captions. Efforts to re-caption web-scraped image datasets for better text fidelity are evident in (Betker et al. 2023; Segalis et al. 2023). Moreover, similar to finetuning LLMs with generated data (Betker et al. 2023; Segalis et al. 2023), DreamSync (Sun et al. 2023) improves T2I synthesis with feedback from vision-language image understanding models, aligning images with textual input and the aesthetic quality of the generated images.

Model-Level Enhancements: At the model level, enhancing the architecture with additional components like attention modules (Feng et al. 2023) offers a training-free solution to enhance model compliance with desired outputs. StructureDiffusion (Feng et al. 2023) and SynGen (Rassin et al. 2023) also work on training-free methods that focus on model's inference time adjustments.

Training-Level Strategies: The integration of supervised finetuning (SFT) with advanced alignment stages, such as reinforcement learning approaches like DPOK (Fan et al. 2024), DDPO (Black et al. 2023), and DPO (Rafailov et al. 2023), shows significant potential in aligning image quality with human preferences. Among these, Diffusion-DPO emerges as an RL-free method, distinct from other RL-based alignment strategies, effectively enhancing human appeal while ensuring distributional integrity (Wallace et al. 2023).

A common drawback of these approaches is the expensive finetuning dataset, as most of them rely on human-annotated data and human evaluation. This paradigm cannot support training an extensive and scalable diffusion model.

2.2 Image Quality Evaluation Methods

Evaluating image quality in a comprehensive manner is pivotal, integrating both automated benchmarks and human assessments to ensure fidelity and aesthetic appeal. The introduction of TIFA (Hu et al. 2023) utilize Visual Question Answering (VQA) models to measure the faithfulness of generated images to text prompts, setting a foundation for subsequent innovations. The CLIP score (Hessel et al. 2021) builds upon CLIP (Radford et al. 2021) enables a reference-free evaluation of image-caption compatibility through the computation of cosine similarity between image and text embeddings, showcasing high correlation with human judgments without needing reference captions. PickScore (Kirstain et al. 2024) leverages user preferences to predict the appeal of generated images, combining CLIP model elements with InstructGPT's reward model objectives (Ouyang et al. 2022) for a nuanced understanding of user satisfaction. Alongside, the aesthetic score (Ke et al. 2023) assesses images based on aesthetics learned from image-comment pairs, providing a richer evaluation that includes composition, color, and style.

3 Methodology

The overall pipeline of AGFSync is illustrated in Fig. 1.

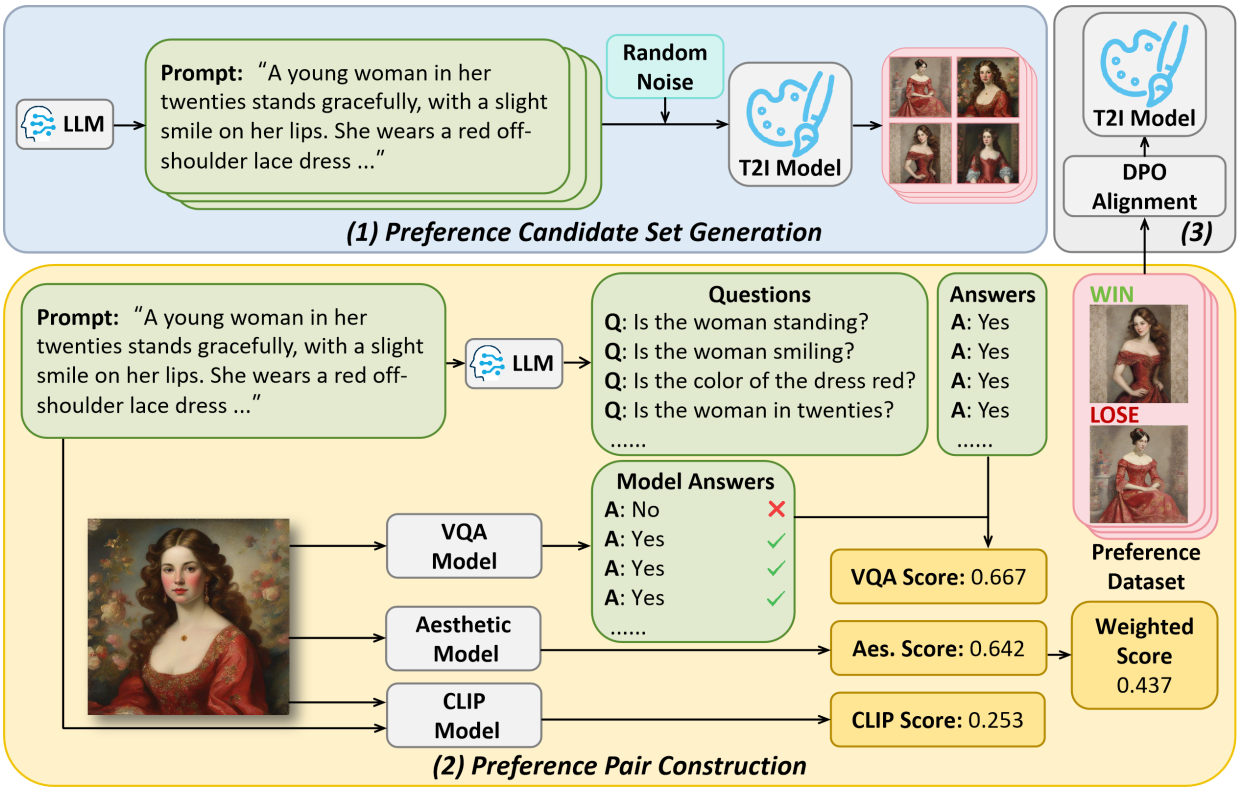


Figure 1: Overall pipeline of AGFSync, which mainly encompasses 3 steps. AGFSync learns from AI-generated feedback data with DPO. AGFSync requires no human annotation, model architecture changes, or reinforcement learning.

3.1 Preference Candidate Set Generation

To encourage the diffusion model to generate diverse style images for further text-image pair preference datasets, we employ LLM to generate prompts $\boldsymbol{c}$ from the instruction that would further feed into the T2I diffusion model serving as image captions. We encourage the LLM to generate 12 categories: Natural Landscapes, Cities and Architecture, People, Animals, Plants, Food and Beverages, Sports and Fitness, Art and Culture, Technology and Industry, Everyday Objects, Transportation, and Abstract and Conceptual Art.

For each category, we utilize in-context learning strategy – carefully craft 5 high-quality examples aimed at guiding the large language model to grasp the core characteristics and contexts of each category, thereby generating new prompts with relevant themes and rich content. Additionally, we emphasize the diversity in prompt lengths, aiming to produce both succinct and elaborate prompts to cater to different generational needs and usage scenarios.

To construct the preference candidate set, we consider a text-conditioned generative diffusion model G for candidate the image generation, where G accept input parameters: text condition $\boldsymbol{c}$ and latent space noise $\boldsymbol{z}_0$. We let the diffusion model to generate N candidate images. To enhance the diversity and distinctiveness of the images produced by the model, we incorporate Gaussian noise into the conditional input $\boldsymbol{c}$ and generate $\boldsymbol{z}_0$ with different random seeds. This approach aims to introduce more randomness and variation to avoid overly uniform or similar generated images. Specifically, the process of generating backup images can be represented as in Eq. (1):

$$\boldsymbol{x}_0 = \mathrm{G}(\boldsymbol{c} + \boldsymbol{n}, \boldsymbol{z}_0) \tag{1}$$

where Gaussian noise $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ is added to the conditional input, increasing the diversity of images.

In practice, by adjusting the value of variance σ and using different random seeds to generate $\boldsymbol{z}_0$, the diversity of the generated images can be controlled. A larger σ value will lead to greater variability in the conditility of the input, but potentially producing more diverse images but might also decrease the relevance of the image to the condition.

Therefore, we currently have the sample $\boldsymbol{c}$ and its corresponding N preference candidate generated images. Next, we will filter and refine these candidates to construct the final preference pair dataset.

3.2 Preference Pair Construction

VQA Questions Generation We also employ the LLM to refine the prompts generated for T2I generation into a series of question-and-answer pairs (QA pairs). By letting Visual Question Answering (VQA) model to answer these questions based on the generated images, the VQA score is calculated. We will establish the preference pair according to multiple image quality scores later.

To make the score easier to calculate, we ensure that the answers to these questions are uniformly "yes" in the in-

struction prompt. To refine the questions, we let the LLM to validate the questions if they are ambiguous or unrelated to the captions, therefore all questions are generated not valid or closely related to the text for answering by the validation process in the instruction prompt.

VQA Score The VQA score is computed by evaluating the correctness of answers provided by the VQA model to the questions generated from the text prompt $\boldsymbol{c}$. For each text prompt $\boldsymbol{c}$, the set of QA pairs is denoted as $\{(Q_i(\boldsymbol{c}), A_i(\boldsymbol{c}))\}$ for $i = 1, \ldots, N_{\boldsymbol{c}}$, where $N_{\boldsymbol{c}}$ is the total number of QA pairs generated for the text prompt $\boldsymbol{c}$, and $\boldsymbol{x}_0$ represents the image generated from $\boldsymbol{c}$.

The VQA model Φ is employed to answer all questions $Q_i(\boldsymbol{c})_{i=1}^{N_{\boldsymbol{c}}}$ based on the image $\boldsymbol{x}_0$. The correctness of the VQA model's answers is evaluated by comparing them to the correct answers $A_i(\boldsymbol{c})$. The VQA score (Hu et al. 2023), which quantifies the consistency between the text and the generated image, is calculated in Eq. (2):

$$s_{\text{VQA}} = \frac{1}{N_{\boldsymbol{c}}} \sum_{i=1}^{N_{\boldsymbol{c}}} \begin{cases} 1 & \text{if } \Phi(\boldsymbol{x}_0, Q_i(\boldsymbol{c})) = A_i(\boldsymbol{c}), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here, the case structure explicitly represents the indicator function, which is 1 if the VQA model's answer matches the correct answer $A_i(\boldsymbol{c})$, and 0 otherwise.

CLIP Score Utilizing the CLIP (Radford et al. 2021) model, we convert the prompt words and the generated image into vector representations, denoted as $\boldsymbol{c}^{(emb)}$ for text and $\boldsymbol{x}_0'$ for the image. The cosine similarity between the two vectors, computed in a shared embedding space, quantifies the alignment between the text and the image, embodying the CLIP Score (Hessel et al. 2021), defined in Eq. (3):

$$s_{\text{CLIP}} = \cos(\boldsymbol{c}^{(emb)}, \boldsymbol{x}_0') = \left(\frac{\boldsymbol{c}^{(emb)}}{||\boldsymbol{c}^{(emb)}||_2} \cdot \frac{\boldsymbol{x}_0'}{||\boldsymbol{x}_0'||_2} \right) * \gamma \tag{3}$$

Aesthetic Score The aesthetic score assesses an image's visual appeal by analyzing multifaceted elements like composition, color harmony, style, and high-level semantics, which collectively contribute to the aesthetic quality of an image (Ke et al. 2023). The evaluation is defined in Eq. (4):

$$s_{\text{Aesthetic}} = AestheticModel(\boldsymbol{x}_0) \tag{4}$$

where $\boldsymbol{x}_0$ signifies the input image, and $AestheticModel(\cdot)$ refers to a sophisticated model function that yields a score reflecting the image's aesthetic appeal on a normalized scale. Higher scores denote a greater aesthetic appeal.

Weighted Score Calculation Consider a set of scores $\{s_1, s_2, \ldots, s_n\}$, where each score s_i corresponds to a distinct evaluation metric utilized. Alongside these scores, let there be a set of weights $W = \{w_1, w_2, \ldots, w_n\}$, with each weight w_i specifically assigned to modulate the influence of its corresponding score s_i.

The composite score for an image $\boldsymbol{x}_0$, which integrates these diverse evaluation metrics, is determined by calculating the sum of the weighted scores. The formula for computing this aggregated score is given by Eq. (5):

$$S(\boldsymbol{x}_0) = \sum_{i=1}^{n} w_i s_i(\boldsymbol{x}_0) \tag{5}$$

where n represents the total number of individual scores. The weighted sum approach facilitates the model's capability to assess images across varied criteria, offering a comprehensive understanding of the image's quality and relevance.

Preference Pair Dataset Construction With the generated set of N images $\boldsymbol{X}_0 = \{\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \ldots, \boldsymbol{x}_0^N\}$ for a given textual prompt $\boldsymbol{c}$, each candidate image is then evaluated to assign the score calculated in multiple aspects as the aforementioned weighted score. To identify the most and least preferred images, which termed as the "winner" and "loser", we apply the selection criteria in Eq. (6) and Eq. (7):

$$\boldsymbol{x}_0^w = \arg \max_{\boldsymbol{x}_0^i \in \boldsymbol{X}_0} S(\boldsymbol{x}_0^i) \tag{6}$$

$$\boldsymbol{x}_0^l = \arg \min_{\boldsymbol{x}_0^i \in \boldsymbol{X}_0} S(\boldsymbol{x}_0^i) \tag{7}$$

This approach yields a preference pair for each textual prompt $\boldsymbol{c}$, represented as $(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)$. The rationale behind selecting the highest and lowest scored images is to capture the widest possible discrepancy in quality and relevance, providing a clear contrast suitable for finetuning with DPO.

3.3 DPO Alignment

Derive from Diffusion-DPO (Rafailov et al. 2023), we consider the preference dataset, denoted as $\mathcal{D} = \{(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)\}$. Applying DPO for diffusion models is modeled as the following objective function $L(\theta)$ in Eq. (8). For the detailed notation of algorithms Eq. (8), please refer to Diffusion-DPO (Rafailov et al. 2023) and DPO (Rafailov et al. 2023).

$$\begin{aligned} L(\theta) = -\mathbb{E}_{(\boldsymbol{x}_0^w, \boldsymbol{x}_0^l) \sim \mathcal{D}, t \sim \mathcal{U}(0,T), \boldsymbol{x}_t^w \sim q(\boldsymbol{x}_t^w | \boldsymbol{x}_0^w), \boldsymbol{x}_t^l \sim q(\boldsymbol{x}_t^l | \boldsymbol{x}_0^l)} \\ \log \sigma(-\beta T \omega(\lambda_t)(\|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w, t, \boldsymbol{c})\|_2^2 \\ - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w, t, \boldsymbol{c})\|_2^2 - (\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l, t, \boldsymbol{c})\|_2^2 \\ - \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, \boldsymbol{c})\|_2^2))) \end{aligned} \tag{8}$$

where $\boldsymbol{x}_t^* = \alpha_t \boldsymbol{x}_0^* + \sigma_t \boldsymbol{\epsilon}^*$, $\boldsymbol{\epsilon}^* \sim \mathcal{N}(0, \boldsymbol{I})$. Here, α_t and σ_t are the noise scheduling functions as defined in (Rombach et al. 2022a). Consequently, $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I})$. Similar to (Wallace et al. 2023), we incorporate T and $\omega(\lambda_t)$ into the constant β.

4 Experimental Setups

4.1 Datasets

To evaluate whether our AGFSync can enhance the performance of text-to-image models across a wide range of prompts, we consider the following benchmarks: **(1) TIFA** (Hu et al. 2023): Based on the correct answers to a series of predefined questions. TIFA employs visual question answering (VQA) models to determine whether the content of generated images accurately reflects the details of the input text. The benchmark itself is comprehensive, encompassing 4,000 different text prompts and 25,000 questions across 12 distinct categories. **(2) HPS v2** (Wu et al. 2023): Human Preference Score v2 (HPS v2) is a benchmark designed to evaluate models' capabilities across a variety of image types. It comprises 3,200 distinct image captions and covers five categories of image descriptions: anime, photo, drawbench, concept-art, and paintings.

4.2 Hyperparameters

For each given text prompt $\boldsymbol{c}$, we let the diffusion model generate $N = 8$ samples as backup images for preference dataset construction. In this process, we add Gaussian noise $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ to the text embedding, where σ is set to 0.1. In the calculation of CLIP score, γ is set to 100, which leads to the CLIP Score range between 0 and 100. We also rescale the VQA score and aesthetic score to $0-100$ by multiplying the original score by 100.

The weighting of each score measurements is allocated as: $w_{\text{VQA}} = 0.35$, $w_{\text{CLIP}} = 0.55$, $w_{\text{Aesthetic}} = 0.1$. Thus, the weighted score S for an image $\boldsymbol{x}$ is calculated as Eq. (9):

$$S(x) = 0.35 s_{\text{VQA}}(x) + 0.55 s_{\text{CLIP}}(x) + 0.1 s_{\text{Aes.}}(x) \quad (9)$$

During the DPO alignment stage, we finetune the original diffusion model. For the SD v1.4 and SD v1.5 models, the learning rate is 5e-7, the batch size is 128, the output image size is 512×512. For the SDXL-base model, the learning rate is 1e-6, the batch size is 64, the output image size is 1024×1024. We finetune the diffusion model for 1,000 steps. The random seed is set to 200 in Fig. 3b and Fig. 3a.

4.3 Baseline Models and Utilized Models

We evaluate AGFSync using Stable Diffusion v1.4 (SD v1.4), Stable Diffusion v1.5 (SD v1.5) (Rombach et al. 2022b), and Stable Diffusion XL Base 1.0 (SDXL-base) (Podell et al. 2023), widely acknowledged in related research as the current leading open-source text-to-image (T2I) models. For prompt construction, we employ ChatGPT (GPT-3.5) (OpenAI 2023). For generating Q&A pairs, we use Gemini Pro (Pichai and Hassabis 2023). Both are accessed through their official API. In addition, we adopt *Salesforce/blip2-flan-t5-xxl* for VQA scoring model (Li et al. 2023), *openai/clip-vit-base-patch16* for evaluating CLIP score (Hessel et al. 2021), *Vila* (Ke et al. 2023) for calculating aesthetic score (Ke et al. 2023), which are consistent with the baseline methods' settings. We also employ GPT-4 Vision (GPT-4V) (Achiam et al. 2023) to simulate human preferences when evaluating the image quality.

5 Experimental Results

5.1 Benchmarking Results on HPS v2

As in Table 1, we test the win rates of models finetuned with our method AGFSync against the original models on the CLIP score and aesthetic score in the HPS v2 benchmark. The experimental results show that with AGFSync, models consistently achieve win rates exceeding 50% across all image categories and both evaluation metrics, CLIP score and aesthetic score, compared to the original baseline SD v1.4, SD v1.5, and SDXL-base models without finetuning. Notably, after AGFSync finetuning, the SDXL-base model not only achieves a win rate of 60.5% in the CLIP score compared to the original model in the anime category images, but also achieves a win rate of 77.4% in the aesthetic score for the same category images. The average win rate of the CLIP score and aesthetic score for the three models increases to 57.2% and 66.7%, respectively, compared to base ones.

Category	SD v1.4		SD v1.5		SDXL-base	
	CLIP	Aes.	CLIP	Aes.	CLIP	Aes.
Anime	57.0%	58.0%	58.6%	66.4%	60.5%	77.4%
Concept-art	55.9%	57.5%	59.3%	65.6%	58.1%	78.3%
Drawbench	56.0%	57.0%	59.0%	57.6%	55.5%	76.6%
Paintings	57.1%	62.0%	56.6%	67.1%	56.6%	74.7%
Photo	55.1%	61.6%	56.9%	60.9%	56.2%	79.3%

Table 1: Comparison of the win rates of SD v1.4, SD v1.5 and SDXL-base with or without our AGFSync on HPS v2.

Evaluate by GPT-4 Vision to Simulate Human Preference In this study, we explore the efficacy of AGFSync in enhancing image generation models, leveraging the capabilities of GPT-4 Vision (GPT-4V) as reported by OpenAI in 2023 (Achiam et al. 2023) to simulate human preferences. Our methodology involves: (1) a comparative analysis of images generated by various diffusion models before and after the application of AGFSync; and (2) a comparative analysis of images generated by the SD v1.4 model after applying AGFSync or other alignment methods. These images, accompanied by their respective descriptions, are submitted to GPT-4V for evaluation based on three critical aspects: **General Preference (Q1)**: "Which image do you prefer?"; **Prompt Alignment (Q2)**: "Which image better fits the text description?"; **Visual Appeal (Q3)**: "Disregarding the prompt, which image is more visually appealing?".

Test Model	Method	General	Faithful	Aesthetic
SD v1.4	vs. Original	62%	58%	65%
	vs. DDPO	68%	78%	82%
	vs. Structured Diffusion	64%	70%	79%
	vs. SynGen	61%	58%	58%
SD v1.5	vs. Original	68%	67%	65%
SDXL-base	vs. Original	62%	69%	76%

Table 2: Win rate results of using GPT-4V to evaluate our finetuned models based on SD v1.4, SD v1.5, and SDXL-base, compared to the original models and models aligned with DDPO, Structured Diffusion, and SynGen (only on SD v1.4), for general preference (Q1), prompt alignment (Q2), and visual appeal (Q3) on the HPS v2 dataset.

The evaluation process involves collecting and analyzing the frequency with which images produced by both the original and the finetuned model are favored under each question category. The results of the comparative analysis of images generated by various diffusion models before and after the application of AGFSync are presented in Table 2, which sequentially displays the performance metrics. The performance reveals that adding AGFSync yields substantial enhancements across all models concerning Q1, Q2, and Q3. Notably, with our AGFSync applied, we achieve an average of 62%, 67%, and 69% win rates across the three aspects

for the SD v1.4, SD v1.5, and SDXL-base models respectively. The results of the comparative analysis of images generated by the SD v1.4 model after applying AGFSync or other alignment methods are also presented in Table 2. The performance shows AGFSync consistently outperforms other baselines (DDPO, Structured Diffusion, and SynGen) applied to the SD v1.4 in all dimensions, achieving average win rates of 64.2%, 68.4%, and 72.8% respectively, when evaluated by GPT-4V on images generated from the HPS v2. These results demonstrate the effectiveness of AGFSync in enhancing performance under various prompts.

Human Evaluation of GPT-4V Judgments To validate the efficacy of GPT-4V for image evaluation and address potential biases in AI assessments, we compare its consistency with human evaluations. We randomly select 9 pairs of images generated by the AGFSync that are favored by GPT-4V. A total of 58 graduate students from China participate in the evaluation. Each participant assesses each image pair based on the criteria Q1, Q2, Q3. Each image pair is independently rated on these three dimensions, resulting in 27 questions per participant (9 image pairs $\times$ 3 dimensions). For Q1, there is 78% agreement between GPT-4V and human evaluations, for Q2, 83%, and for Q3, 70%. All dimensions show agreement rates above 50%, indicating that GPT-4V's evaluations align closely with human preferences and confirming its reliability as a tool for reducing individual biases and maintaining objectivity in image evaluation.

5.2 Benchmarking Results on TIFA

In Table 3, we further test our method on the TIFA benchmark, highlighting AGFSync's SOTA performance on VQA score and aesthetic score over other latest SOTA alignment methods. Specifically, we compare three types of alignment methods: training-free approaches capable of modifying outputs without retraining the model, such as StructureDiffusion (Feng et al. 2023) and SynGen (Rassin et al. 2023); reinforcement learning (RL)-based methods aimed at improving model outputs, such as DPOK (Fan et al. 2024) and DDPO (Black et al. 2023); and methods like DreamSync (Sun et al. 2023), which employ self-training strategy but focus on SFT stage. Given that these baseline methods are all based on SD v1.4, we ensure a fair comparison by using the same version of the SD model as the foundation and employing the same VQA model (BLIP-2) for evaluation. Results reveal that our method AGFSync can simultaneously improve the text fidelity and visual quality of SD v1.4, SD v1.5, and SDXL-base models. For SD v1.4, AGFSync achieves an improvement of 1.3% of VQA score and 3.3% of aesthetic score, with a total improvement of 4.6% on the TIFA benchmark, higher than all baseline models. Note that although DPOK shows a 1.9% improvement on aesthetic score, it reduces the model's text faithfulness through VQA score. For SD v1.5 and SDXL-base, our method AGFSync leads to improvements of 1.6% and 1.1% for SD v1.5, 1.3% and 4.3% for SDXL-base on VQA score and aesthetic score respectively, which are both higher than the results achieved by DreamSync finetuned using self-training SFT.

Model		Alignment	s_{VQA}	$s_{\text{Aes.}}$	Sum
SD v1.4		No alignment	76.6	44.6	-
	Training-Free	SynGen	76.8 (+0.2)	42.4 (−2.2)	−2.0
		StructureDiffusion	76.5 (−0.1)	41.5 (−3.1)	−3.0
	RL	DPOK	76.4 (−0.2)	46.5 (+1.9)	+1.7
		DDPO	76.7 (+0.1)	43.5 (−1.1)	−1.0
	Self-Training	DreamSync	77.6 (+1.0)	44.9 (+0.3)	+1.3
		AGFSync (Ours)	**77.9** (+1.3)	**47.9** (+3.3)	+4.6
SD v1.5		No alignment	77.1	48.0	-
		DreamSync	77.7 (+0.6)	47.6 (−0.4)	+0.2
		AGFSync (Ours)	**78.7** (+1.6)	**49.1** (+1.1)	+2.7
SDXL-base		No alignment	82.0	60.9	-
		DreamSync	83.1 (+1.1)	64.1 (+3.2)	+4.3
		AGFSync (Ours)	**83.3** (+1.3)	**65.2** (+4.3)	+5.5

Table 3: Results of different alignment methods on VQA score and aesthetic score on the TIFA benchmark. The best scores for each model type are in Bold. Column "Sum" denotes the sum of improvements on s_{VQA} and $s_{\text{Aes.}}$.

5.3 Experiment of Comparing the Dataset Quality between MJHQ-30K and AGFSync

MJHQ-30K is a benchmark dataset used for automatically evaluating the aesthetic quality of models (Li et al. 2024). It consists of high-quality images curated from Midjourney, covering 10 common categories, with each category containing 3,000 samples. MJHQ-30K can also serve as a training dataset for general SFT. To compare the quality of the preference dataset built using AGFSync with MJHQ-30K, we finetune SD v1.4, SD v1.5 and SDXL-base using MJHQ-30K and compare their performance against the SD v1.4, SD v1.5 and SDXL-base finetuned with AGFSync. As shown in Table 4, AGFSync applied to SD v1.4, SD v1.5 and SDXL-base achieve superior improvements in text alignment. Although finetuning SD v1.4 and SD v1.5 with the MJHQ-30K dataset results in the highest improvement in aesthetic scores, this is because the images in MJHQ-30K is generated by Midjourney, which have much higher aesthetic quality than those generated by SD v1.4 and SD v1.5 for self-training. When finetuning SDXL-base with MJHQ-30K, the improvement in aesthetic scores is less pronounced compared to AGFSync, demonstrating the effectiveness of the preference dataset constructed using AGFSync.

Model	Alignment	s_{VQA}	$s_{\text{Aes.}}$	Sum
SD v1.4	No alignment	76.6	44.6	-
	MJHQ-30K+SFT	77.6 (+1.0)	**48.3** (+3.7)	**+4.7**
	AGFSync (Ours)	**77.9** (+1.3)	47.9 (+3.3)	+4.6
SD v1.5	No alignment	77.1	48.0	-
	MJHQ-30K+SFT	78.3 (+1.2)	**49.3** (+1.3)	+2.5
	AGFSync (Ours)	**78.7** (+1.6)	49.1 (+1.1)	**+2.7**
SDXL-base	No alignment	82.0	60.9	-
	MJHQ-30K+SFT	82.6 (+0.6)	61.1 (+0.2)	+0.8
	AGFSync (Ours)	**83.3** (+1.3)	**65.2** (+4.3)	**+5.6**

Table 4: SD v1.4, SD v1.5 and SDXL-base's results of general SFT setting on MJHQ-30K compared to AGFSync.

5.4 Experiment of Gaussian Noise for Diversity

To demonstrate how Gaussian noise n added to condition c enhances image diversity, we generate N candidates using the prompt "wild animal" with varying noise weights (Fig. 2). As noise weight increases, we observe greater variety in generated animal species, with maximum diversity achieved at weight 1. This confirms that Gaussian noise effectively broadens the model's exploration space by expanding the conditional input coverage.

Figure 2: Image noise levels (shown on left) and their effect on diversity of generated images.

5.5 Ablation Experiment of Multi-Aspect Scoring

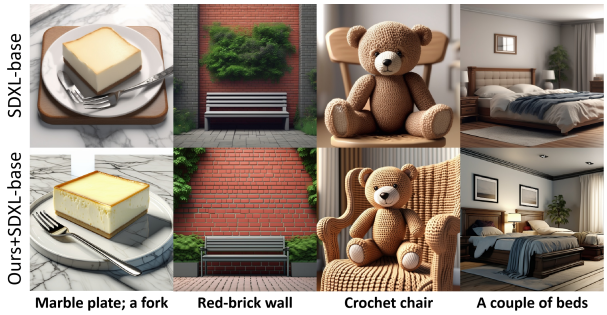
As depicted in Table 5, to validate the efficacy of the three scores that we employ for image quality assessment, we conduct a thorough ablation study. We train the SD v1.5 model on preference datasets constructed with different combinations of the three scores, along with PickScore (Kirstain et al. 2024). As in AGFSync, with training model on preference datasets built using a combination of CLIP score, VQA score, and aesthetic score result in the greatest improvement across all three metrics. While other combinations often show a decrease in certain metrics rather than a consistent improvement on all metrics.

Applied Measures				s_{CLIP}	s_{VQA}	$s_{\text{Aes.}}$
+CLIP	+VQA	+Aes.	+Pick			
-	-	-	-	27.0	77.1	48.0
✓	-	-	-	27.2	77.7	47.3
-	✓	-	-	27.1	77.4	45.7
-	-	✓	-	27.0	76.8	48.6
✓	✓	-	-	27.2	77.5	47.0
✓	-	✓	-	27.2	77.8	47.2
-	✓	✓	-	27.1	77.2	48.2
-	-	-	✓	27.1	78.0	47.8
✓	✓	✓	-	**27.3**	**78.7**	**49.1**

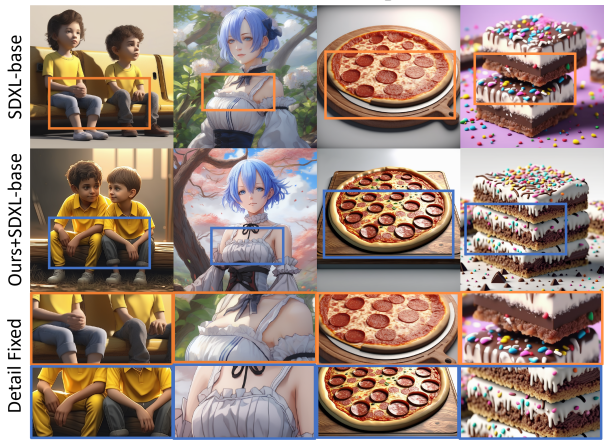
Table 5: Results of applied different scoring measures. Experiments are conducted with SD v1.5 on TIFA.

5.6 Comparing Visual Faithfulness & Coherence

Figs. 3a and 3b compare images generated by SDXL-base and AGFSync using identical prompts. While SDXL-base generates vivid images, they sometimes deviate from input descriptions (Fig. 3a) or contain unrealistic details (Fig. 3b). For example, SDXL-base produces unnatural wrinkles on a girl's chest and physically impossible floating cakes. AGFSync improves the consistency of generated images with prompts and enhances adherence to real-world physics.



(a) Text faithfulness comparison



(b) Adherence to real-world rules

Figure 3: Comparison between SDXL-base and AGFSync (Ours)+SDXL-base. (a) Bold text indicates discrepancies between the output images of SDXL-base with input prompts. (b) Third row compares details, showing AGFSync's improved coherence and detail.

6 Conclusions

This paper introduces a text-to-image generation framework AGFSync. By leveraging Direct Preference Optimization (DPO) and multi-aspect AI feedback, AGFSync significantly enhances the prompt following ability and image quality regarding style, coherence, and aesthetics. Extensive experiments on the HPS v2 and TIFA benchmark demonstrate that AGFSync outperforms baseline models in terms of VQA scores, CLIP score, aesthetic evaluation. Based on an AI-driven feedback loop, AGFSync eliminates the need for costly human-annotated data and manual intervention, paving the way for scalable alignment techniques.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2023YFB3309000), the National Natural Science Foundation of China under Grants U2241217, 62473027 and 62473029.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Dai, X.; Hou, J.; Ma, C.-Y.; Tsai, S.; Wang, J.; Wang, R.; Zhang, P.; Vandenhende, S.; Wang, X.; Dubey, A.; Yu, M.; Kadian, A.; Radenovic, F.; Mahajan, D.; Li, K.; Zhao, Y.; Petrovic, V.; Singh, M. K.; Motwani, S.; Wen, Y.; Song, Y.; Sumbaly, R.; Ramanathan, V.; He, Z.; Vajda, P.; and Parikh, D. 2023. Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack. arXiv:2309.15807.

Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; and Lee, K. 2024. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. arXiv:2212.05032.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Hu, Y.; Liu, B.; Kasai, J.; Wang, Y.; Ostendorf, M.; Krishna, R.; and Smith, N. A. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*.

Ke, J.; Ye, K.; Yu, J.; Wu, Y.; Milanfar, P.; and Yang, F. 2023. VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10041–10051.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2024. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Li, D.; Kamko, A.; Akhgari, E.; Sabet, A.; Xu, L.; and Doshi, S. 2024. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation. arXiv:2402.17245.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

OpenAI. 2023. Introducing ChatGPT.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pichai, S.; and Hassabis, D. 2023. Introducing Gemini: our largest and most capable AI model. *Google. Retrieved December*, 8: 2023.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Qin, Y.; Shi, Z.; Yu, J.; Wang, X.; Zhou, E.; Li, L.; Yin, Z.; Liu, X.; Sheng, L.; Shao, J.; et al. 2024. WorldSimBench: Towards Video Generation Models as World Simulators. *arXiv preprint arXiv:2410.18072*.

Qin, Y.; Zhou, E.; Liu, Q.; Yin, Z.; Sheng, L.; Zhang, R.; Qiao, Y.; and Shao, J. 2023. Mp5: A multi-modal open-ended embodied system in minecraft via active perception. *arXiv preprint arXiv:2312.07472*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Rassin, R.; Hirsch, E.; Glickman, D.; Ravfogel, S.; Goldberg, Y.; and Chechik, G. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schuhmann, C. 2022. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/. Accessed: 2024-02-29.

Segalis, E.; Valevski, D.; Lumen, D.; Matias, Y.; and Leviathan, Y. 2023. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*.

Sun, J.; Fu, D.; Hu, Y.; Wang, S.; Rassin, R.; Juan, D.-C.; Alon, D.; Herrmann, C.; van Steenkiste, S.; Krishna,

R.; and Rashtchian, C. 2023. DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback. arXiv:2311.17946.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2023. Diffusion Model Alignment Using Direct Preference Optimization. arXiv:2311.12908.

Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341*.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2024a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zhou, E.; Qin, Y.; Yin, Z.; Huang, Y.; Zhang, R.; Sheng, L.; Qiao, Y.; and Shao, J. 2024b. MineDreamer: Learning to Follow Instructions via Chain-of-Imagination for Simulated-World Control. *arXiv preprint arXiv:2403.12037*.

Zhou, E.; Su, Q.; Chi, C.; Zhang, Z.; Wang, Z.; Huang, T.; Sheng, L.; and Wang, H. 2024c. Code-as-Monitor: Constraint-aware Visual Programming for Reactive and Proactive Robotic Failure Detection. *arXiv preprint arXiv:2412.04455*.