UniFL: Improve Stable Diffusion via Unified Feedback Learning

Jiacheng Zhang[1,2*], Jie Wu[1*†], Yuxi Ren[1], Xin Xia[1], Huafeng Kuang[1],
Pan Xie[1], Jiashi Li[1], Xuefeng Xiao[1], Min Zheng[1], Lean Fu[1], and Guanbin Li[2]

[1]ByteDance Inc., China, [2]Sun Yat-sen University, China
https://uni-fl.github.io/

Fig. 1: We propose ***UniFL***, a unified framework that leverages feedback learning to elevate the visual generation quality, enhance preference aesthetics, and accelerate the inference process. The figure illustrates the outcomes obtained by optimizing SDXL through UniFL, and *the last three images depict the results with 4 steps inference.*

Abstract. Diffusion models have revolutionized the field of image generation, leading to the proliferation of high-quality models and diverse downstream applications. However, despite these significant advancements, the current competitive solutions still suffer from several limitations, including inferior visual quality, a lack of aesthetic appeal, and inefficient inference, without a comprehensive solution in sight. To address these challenges, we present ***UniFL***, a unified framework that leverages

---

*Equal Contribution.
†Project Lead.

feedback learning to enhance diffusion models comprehensively. UniFL stands out as a universal, effective, and generalizable solution applicable to various diffusion models, such as SD1.5 and SDXL. Notably, UniFL incorporates three key components: perceptual feedback learning, which enhances visual quality; decoupled feedback learning, which improves aesthetic appeal; and adversarial feedback learning, which optimizes inference speed. In-depth experiments and extensive user studies validate the superior performance of our proposed method in enhancing both the quality of generated models and their acceleration. For instance, UniFL surpasses ImageReward by **17%** user preference in terms of generation quality and outperforms LCM and SDXL Turbo by **57%** and **20%** in 4-step inference. Moreover, we have verified the efficacy of our approach in downstream tasks, including Lora, ControlNet, and AnimateDiff.

Keywords: Diffusion Models · Feedback Learning · Acceleration

1   Introduction

The emergence of diffusion models has catapulted the field of Text-to-Image (T2I) into a realm of unparalleled progress, marked by notable contributions like DALLE-3 [34], Imagen [38], Midjourney [49], and etc,. In particular, the introduction of open-source image generation models, exemplified by stable diffusion [35], has inaugurated a transformative era of Text-to-Image, giving rise to numerous downstream applications such as T2I Personalization [15, 21, 37, 60], Controllable Generation [29, 33, 61] and Text-to-Video (T2V) Generation [16, 18, 53]. Despite the remarkable progress achieved thus far, current stable diffusion-based image generation models still exhibit certain limitations. i) *Inferior Quality*: the generated images often suffer from poor quality and lack authenticity. Examples include characters with incomplete limbs or distorted body parts, as well as limited fidelity in terms of style representation. ii) *Lack Aesthetics*: there is a notable bias in the aesthetic appeal of the generated images, as they often fail to align with human preferences. Deficiencies in crucial aspects such as details, lighting, and atmosphere further contribute to this aesthetic disparity. iii) *Inference Inefficiency*: the iterative denoising process employed by diffusion models introduces inefficiencies that significantly impede inference speed, thereby limiting the practicality of these models in various application scenarios.

Recently, numerous works have endeavored to address the aforementioned challenges. For instance, SDXL [32] enhances the generation quality of diffusion models by refining training strategies, while RAPHAEL [59] resorts to the techniques of Mixture of Experts (MoE) [14, 44, 63]. RAFT [11], HPS [54, 55], ImageReward [57], and DPO [50] propose incorporating human feedback to guide diffusion models toward aligning with human preferences. SDXL Turbo [40], PGD [39], and LCM [27, 28], on the other hand, tackle the issue of inference acceleration through techniques like distillation and consistency models [46]. However, these methods primarily concentrate on tackling individual problems through specialized designs, which poses a significant challenge for the direct

integration of these techniques. For example, MoE significantly complicates the pipeline, making the acceleration method infeasible, and the consistency models [46] alter the denoising process of the diffusion model, making it arduous to directly apply the ReFL framework proposed by ImageReward [57]. Naturally, a pertinent question arises: *Can we devise a more effective approach that comprehensively enhances diffusion models in terms of image quality, aesthetic appearance, and generation speed?*

In this paper, we present UniFL, an innovative approach that offers a comprehensive improvement to diffusion models through unified feedback learning. UniFL aims to elevate the visual generation quality, enhance preference aesthetics, and accelerate the inference process. To achieve these objectives, we propose three key components. Firstly, we introduce a pioneering perceptual feedback learning (PeFL) framework that effectively harnesses the extensive knowledge embedded within diverse existing perceptual models to improve visual generation quality. This framework enables the provision of more precise and targeted feedback signals, ultimately enhancing the quality of visual generation in various aspects.

Secondly, we employ decoupled feedback learning to optimize aesthetic quality. By breaking down the coarse aesthetic concept into distinct aspects such as color, atmosphere, and texture, UniFL simplifies the challenge of aesthetic optimization. Additionally, we introduce an active prompt selection strategy to choose the prompts that are more informative and diverse to facilitate more efficient aesthetics preference feedback learning.

Lastly, UniFL develops adversarial feedback learning, wherein the reward model and diffusion model are trained adversarially, enabling the samples under the low denoising steps to be well optimized via the reward feedback, and finally achieves superior inference acceleration. UniFL presents a unified formulation of feedback learning that is both straightforward and versatile, making it adaptable to a wide range of models and yielding impressive improvements. Our contributions are summarized as follows:

- **New Insight**: Our proposed method, UniFL, introduces a unified framework of feedback learning to optimize the visual quality, aesthetics, and inference speed of diffusion models. To the best of our knowledge, UniFL offers the first attempt to address both generation quality and speed simultaneously, offering a fresh perspective in the field.
- **Novelty and Pioneering**: In our work, we shed light on the untapped potential of leveraging existing perceptual models in feedback learning for diffusion models. We highlight the significance of decoupled reward models and elucidate the underlying acceleration mechanism through adversarial training. We believe our ablation experiments provide valuable insights that enrich the community's understanding of these techniques.
- **High Effectiveness**: Through extensive experiments, we demonstrate the substantial improvements achieved by UniFL across multiple types of diffusion models, including SD1.5 and SDXL, in terms of generation quality

and acceleration. Furthermore, UniFL outperforms competitive existing approaches and exhibits strong generalization on various downstream tasks.

2   Related Work

2.1   Text-to-Image Diffusion Models

Recently, diffusion models have gained substantial attention and emerged as the de facto method for text-to-image generation, surpassing traditional probabilistic models like GAN [17] and VAE [22]. Numerous related works have been proposed, including GLIDE [30], DALL-E2 [34], Imagen [38], CogView [10] and many others. Among these, Latent Diffusion Models (LDM) [35] extend the diffusion process to the latent space and significantly improve the training and inference efficiency of the diffusion models, opening the door to diverse applications such as controllable generation [33, 61], image editing [3, 19, 29], and image personalization [15, 21, 37] and so on. Despite the progress achieved thus far, current text-to-image diffusion models still have limitations in *inferior visual generation quality, deviations from human aesthetic preferences, and inefficient inference.* The target of this work is to offer a comprehensive solution that effectively addresses these issues.

2.2   Improvements on Text-to-Image Diffusion Models

Given the aforementioned limitations, researchers have proposed various methods to tackle these issues. Notably, [6, 32, 59] focuses on improving generation quality through more advanced training strategies. Building upon the success of reinforcement learning with human feedback (RLHF) [1, 31] in the field of LLM, [2, 54, 55, 57, 64] explore the incorporation of human feedback to improve image aesthetic quality. On the other hand, [27, 28, 39, 41, 46] concentrate on acceleration techniques, such as distillation and consistency models [46] to achieve inference acceleration. While these methods have demonstrated their effectiveness in addressing specific challenges, their independent nature makes it challenging to combine them for comprehensive improvements. In contrast, our study unifies the objective of enhancing visual quality, aligning with human aesthetic preferences, and acceleration through the feedback learning framework.

3   Preliminaries

Text-to-Image Diffusion Model. Text-to-image diffusion models leverage diffusion modeling to generate high-quality images based on textual prompts via the diffusion model, which generates desired data samples from Gaussian noise through a gradual denoising process. During pre-training, a sampled image x is first processed 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 $\alpha_t = 1 - \beta_t$. 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:

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

Reward Feedback Learning. Reward feedback learning(ReFL) [57] 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. These data are then utilized to train a human preference reward model, which serves as an encoding mechanism for capturing human preferences. More specifically, considering two candidate generations, denoted as x_w (preferred generation) and x_l (unpreferred one), the loss function for training 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 feedback data, $\sigma(\cdot)$ represents the sigmoid function, and c corresponds to the text prompt. The reward model r_θ is optimized to produce a preference-aligned score that aligns with human preferences. In the Preference Fine-tuning phase, ReFL begins with an input prompt c, initializing a latent variable x_T at random. The latent variable is then progressively denoised until reaching a randomly selected timestep t. At this point, the denoised image x_0' is directly predicted from x_t. The reward model obtained from the previous phase is applied to this denoised image, generating the expected preference score $r_\theta(c, x_0')$. This preference score enables the fine-tuning of the diffusion model to align more closely with human preferences:

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

Our method follows a similar learning framework to ReFL but devises several novel components to enable comprehensive improvements.

4 UniFL: Unified Feedback Learning

Our proposed method, UniFL, aims to improve the stable diffusion in various aspects, including visual generation quality, human aesthetic quality, and inference efficiency. our method takes a unified feedback learning perspective, offering a comprehensive and streamlined solution.

An overview of UniFL is illustrated in Fig.2. In the following subsections, we delve into the details of three key components: perceptual feedback learning to enhance visual generation quality (Sec. 4.1); decoupled feedback learning to improve aesthetic appeal (Sec. 4.2); and adversarial feedback learning to facilitate inference acceleration (Sec. 4.3).
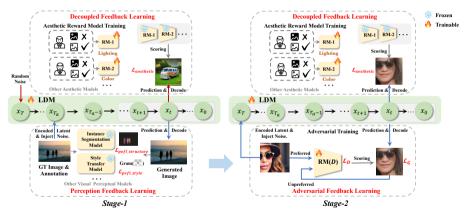
Fig. 2: The overview of **UniFL**, which leverages an unified feedback learning framework to comprehensively enhance the model performance and inference speed. The training process of UniFL is divided into two stages, the first stage aims to improve visual quality and aesthetics, and the second stage speeds up model inference. The gray area within the denoise steps are the timesteps where feedback learning optimizes.

4.1 Perceptual Feedback Learning

Current diffusion models exhibit limitations in achieving high-quality visual generation, particularly in areas such as image style shift and object structure distortion. These limitations stem from the reliance on reconstruction loss solely in the latent space, which lacks supervision based on visual perception in the image space. To address this issue, as illustrated in Fig.3, we propose Perceptual Feedback Learning (PeFL) to fine-tune diffusion model using visual feedback provided by existing perceptual models. Our key insight is that various visual perception models already encapsulate rich visual priors from diverse aspects. The complete PeFL process is summarized in Algorithm 1. In contrast to ReFL, which starts from a randomly initialized latent representation and only considers the text prompt as a condition, PeFL incorporates image content as an additional visual condition for perceptual guidance. Specifically, given a text-image pair, (c, x), we first select a forward step T_a and inject noise into the ground truth image to obtain a conditional latent sequence $x_0 \rightarrow x_1... \rightarrow x_{T_a}$. Subsequently, we randomly select a denoising time step t and denoising from x_{T_a}, yielding $x_{T_a} \rightarrow x_{T_a-1}... \rightarrow x_t$. Next, we directly predict $x_t \rightarrow x_0'$. Various perceptual models are employed to provide feedback on x_0' for specific visual aspects:

i) **Style**: To capture image style, we employ the VGG model to encode image features and extract visual style using the widely adopted gram matrix in style transfer. The feedback on style is calculated as follows:

$$\mathcal{L}(\theta)_{pefl\_style} = \mathbb{E}_{x_0 \sim \mathcal{D}, x_0' \sim G(x_{t_a})} \|Gram(V(x_0')) - Gram(V(x_0))\|_2, \quad (4)$$

where V is the VGG network, and $Gram$ is the calculation of the gram matrix.

ii) **Structure**: For extracting visual structure information, we utilize visual instance segmentation models, as instance masks provide fundamental object

---

Algorithm 1 Perceptual Feedback Learning (PeFL) for LDMs

1: **Dataset:** Perceptual text-image dataset $\mathcal{D} = \{(\text{txt}_1, \text{img}_1), ...(\text{txt}_n, \text{img}_n)\}$
2: **Input:** LDM with pre-trained parameters w_0, visual perceptual model $m_.$, visual perceptual loss function Φ, visual perceptual loss weight λ
3: **Initialization:** The number of noise scheduler time steps T, add noise timestep T_a, denoising time step t.
4: **for** perceptual data point $(\text{txt}_i, \text{img}_i) \in \mathcal{D}$ **do**
5:     $x_0 \leftarrow \text{VaeEnc}(\text{img}_i)$ // Obtain the latent representation of ground truth image
6:     $x_{T_a} \leftarrow \text{AddNoise}(x_0)$ // Add noise into the latent according to Eq.1
7:     **for** $j = T_a, ..., t+1$ **do**
8:         **no grad:** $x_{j-1} \leftarrow \text{LDM}_{w_i}\{x_j\}$
9:     **end for**
10:     **with grad:** $x_{t-1} \leftarrow \text{LDM}_{w_i}\{x_t\}$
11:     $x_0^{'} \leftarrow x_{t-1}$ // Predict the original latent by noise scheduler
12:     $\text{img}_i^{'} \leftarrow \text{VaeDec}(x_0^{'})$ // From latent to image
13:     $\mathcal{L}_{pefl} \leftarrow \lambda\Phi(m(\text{img}_i^{'}), GT(\text{img}_i))$ // PeFL loss by perceptual model
14:     $w_{i+1} \leftarrow w_i$ // Update LDM_{w_i} using PeFL loss
15: **end for**

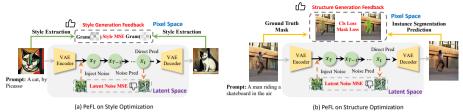(a) PeFL on Style Optimization   (b) PeFL on Structure Optimization

Fig. 3: The illustration of the PeFL on the (a) style and (b) structure optimization. The original noise MSE loss in the latent space only cares about the coarse reconstruction and overlooks the particular visual aspect of the generated image, which can be boosted by the existing various perceptual models via feedback learning.

structure descriptions. The objective is formulated as:

$$\mathcal{L}(\theta)_{pefl\_structure} = \mathbb{E}_{x_0 \sim \mathcal{D}, x_0^{'} \sim G(x_{t_a})} \mathcal{L}_{instance}(m_I(x_0^{'}), GT(x_0)) \qquad (5)$$

where m_I is the instance segmentation model, $GT(x_0)$ is the ground truth instance segmentation annotation and $\mathcal{L}_{instance}$ is the instance segmentation loss.

The flexibility of PeFL allows us to leverage various existing visual perceptual models, for example, semantic segmentation models, to provide specific visual feedback. More experiments and results can be found in the Appendix.

4.2   Decoupled Feedback Learning

Decoupled Aesthetic Fine-tuning. Unlike objective visual quality, aesthetic quality is abstract and subjective, necessitating human aesthetic preference feedback to guide the model toward optimization based on human preferences. ImageReward [57] addresses this issue via a human preference reward model trained on collected preference data within the ReFL framework. While effective, we argue that ImageReward is suboptimal as it relies on a single reward model trained with coarse annotated aesthetic preference data. The primary challenge arises from the attempt to encapsulate human preferences across multiple dimensions

within a single reward model, which will result in inherent conflicts, as evidenced in certain Large Language Model (LLM) studies [48]. To address this problem, we propose decoupling the different aesthetic dimensions during preference modeling to enable more effective aesthetic feedback learning. Specifically, we decompose the general aesthetic concept into representative dimensions and annotate them separately. These dimensions include color, layout, lighting, and detail. The data collection process is described in detail in the Appendix. Subsequently, we train an aesthetic preference reward model using this annotated data according to Eq.2. The objective of the decoupled feedback learning is:

$$\mathcal{L}(\theta)_{aes} = \sum_d \mathbb{E}_{c \sim p(c)} \mathbb{E}_{x'_0 \sim p(x'_0|c)} [\texttt{ReLU}(\alpha_d - r_d(x'_0, c))] \tag{6}$$

r_d is the aesthetic reward model on d dimension, $d \in \{\text{color}, \text{layout}, \text{detail}, \text{lighting}\}$ and α_d are the dimension-aware hinge cofficient for loss calculation.

Active Prompt Selection. We observed that when using randomly selected prompts for preference fine-tuning, the diffusion model tends to rapidly overfit due to the limited semantic richness, leading to diminished effectiveness of the reward model. This phenomenon is commonly referred to as overoptimization [62]. To address this issue, we further propose an active prompt selection strategy, which selects the most informative and diverse prompt from a prompt database. This selection process involves two key components: a semantic-based prompt filter and nearest neighbor prompt compression. By leveraging these techniques, the over-optimization can be greatly mitigated, achieving more efficient aesthetic reward fine-tuning. More details of this strategy are presented in the Appendix.

4.3   Adversarial Feedback Learning

The slow iterative denoising process employed in text-to-image diffusion models poses a significant hindrance to their practical application. To address this limitation, recent advancements, such as UFOGen [58] and SDXL-Turbo [42], have proposed incorporating adversarial training objectives into fine-tuning diffusion models. Building upon this insight, we introduce an adversarial feedback-learning method that combines feedback learning with the adversarial objective, aiming to accelerate the inference process.

The original optimization objective of the diffusion model seeks to increase the reward score of the output image, with the reward model held constant. Rather than freeze the reward model, we incorporate the optimization of an adversarial reward model $r_a(\cdot)$ during the fine-tuning process, treating it as a **discriminator**. In this way, the diffusion model serves as the generator and is optimized to enhance the reward score, while the reward model acts as the discriminator, aiming to distinguish between preferred and unpreferred samples. Consequently, the objective of adversarial feedback learning can be reformulated as follows:

$$\begin{aligned} \mathcal{L}^G(\theta) &= \mathbb{E}_{c \sim p(c)} \mathbb{E}_{x'_0 \sim p(x'_0|c)} [-r_a(x'_0, c)], \\ \mathcal{L}^D(\phi) &= -\mathbb{E}_{(x_0, x'_0, c) \sim \mathcal{D}_{train}, t \sim [1,T]} [\log \sigma(r_a(x_0)) + \log(1 - \sigma(r_a(x'_0)))]. \end{aligned} \tag{7}$$

where θ and ϕ are the parameters of the diffusion model and discriminator. In practice, we follow PeFL to achieve adversarial training, considering the GT image as the preferred sample and the denoised image as the unpreferred sample. In this way, we continually guide the diffusion model to generate samples with higher fidelity and visual quality, which greatly accelerates the inference speed.

4.4   Training Pipeline

Our training process consists of two stages, each targeting a specific objective. In the first stage, we focus on improving visual generation quality and aesthetics. In the second stage, we apply adversarial feedback learning to accelerate the diffusion inference speed, which simultaneously updates the diffusion model and reward model with the adversarial training objective. We also integrate decoupled feedback learning to maintain aesthetics.

$$\mathcal{L}^1(\theta) = \mathcal{L}(\theta)_{pefl} + \mathcal{L}(\theta)_{aes}; \quad \mathcal{L}^2(\theta, \phi) = \mathcal{L}^G(\theta) + \mathcal{L}^D(\phi)) + \mathcal{L}(\theta)_{aes} \quad (8)$$

5   Experiments

5.1   Implementation Details and Metrics

Dataset. For the PeFL training stage, we curated a large and high-quality dataset consisting of approximately 150k artist-style text images for style optimization and utilized COCO2017 [26] train split dataset with instance annotations and captions for structure optimization. Additionally, we collected the human preference dataset for the decoupled aesthetic feedback learning from diverse aspects (such as color, layout, detail, and lighting). The 100,000 prompts are selected for aesthetic optimization from DiffusionDB [52] through active prompt selection. During the adversarial feedback learning, we simply utilize an aesthetic subset of LAION [43] with image aesthetic scores above 5.

Training Setting. We employ the VGG-16 [45] network to extract gram matrix concepts for style PeFL and utilize the SOLO [51] as the instance segmentation model. We utilize the DDIM scheduler with a total of 20 inference steps. $T_a = 10$ and the optimization steps $t \in [5, 0]$ during PeFL training. For adversarial feedback learning, we initialize the adversarial reward model with the weight of the aesthetic preference reward model of details. During adversarial training, the optimization step is set to $t \in [0, 20]$ encompassing the entire process.

Baseline Models. We choose two representative text-to-image diffusion models with distinct generation capacities to comprehensively evaluate the effectiveness of UniFL, including (i) SD1.5 [36]; (ii) SDXL [32]. Based on these models, we pick up several state-of-the-art methods(i.e. ImageReward [57], Dreamshaper [9], and DPO [50] for generation quality improvement, LCM [27], SDXL-Turbo [40], and SDXL-Lightning [25] for inference acceleration) to compare the effectiveness of quality improvement and acceleration. All results of these methods are reimplemented with the official code provided by the authors.

Model	Inference Steps	FID↓	CLIP Score↑	Aesthetic↑
SD1.5	20	37.99	0.308	5.26
SD1.5 ImageReward [57]	20	<u>32.31</u>	0.312	5.37
SD1.5 DreamShaper [9]	20	34.21	<u>0.313</u>	<u>5.44</u>
SD1.5 DPO [50]	20	32.83	0.308	5.22
SD1.5 UniFL	20	**31.14**	**0.318**	**5.54**
SD1.5	4	42.91	0.279	5.16
SD1.5 LCM [27]	4	42.65	<u>0.314</u>	<u>5.71</u>
SD1.5 DreamShaper LCM [28]	4	<u>35.48</u>	0.314	5.58
SD1.5 UniFL	4	**33.54**	**0.316**	**5.88**
SDXL	25	27.92	0.321	5.65
SDXL ImageReward [57]	25	<u>26.71</u>	0.319	<u>5.81</u>
SDXL DreamShaper [9]	25	28.53	0.321	5.65
SDXL DPO [50]	25	35.30	<u>0.325</u>	5.64
SDXL UniFL	25	**25.54**	**0.328**	**5.98**
SDXL	4	125.89	0.256	5.18
SDXL LCM [27]	4	<u>27.23</u>	0.322	5.48
SDXL Turbo [40]	4	30.43	<u>0.325</u>	5.60
SDXL Lighting [25]	4	28.48	0.323	<u>5.66</u>
SDXL UniFL	4	**26.25**	**0.325**	**5.87**

Table 1: The quantitative comparison between our method and other methods on SD1.5 and SDXL architecture. The best performance is highlighted with bold font, and the second-highest performance is shown with underline.

Evaluation Metrics We generate the 5K image with the prompt from the COCO2017 validation split to report the Fréchet Inception Distance(FID) [20] as the overall visual quality metric. We also report the CLIP score with ViT-B-32 [12] and the aesthetic score with LAION aesthetic predictor to evaluate the text-to-image alignment and aesthetic quality of the generated images, respectively. Given the subjective nature of quality evaluations, we also conducted comprehensive user studies to obtain a more accurate evaluation. **For more implementation details of UniFL, please refer to the Appendix.**

5.2   Main Results

Quantitative Comparison. Tab.1 summarize the quantitative comparisons with competitive approaches on SD1.5 and SDXL. Generally, UniFL exhibits consistent performance improvement on both architectures and surpasses the existing methods of focus on improving generation quality or acceleration. Specifically, DreamShaper achieves considerable aesthetic quality in SD1.5(5.44), while ImageReard obtains the best performance in SDXL(5.88). Even though, UniFL surpasses these methods on all of these metrics in both SD1.5 and SDXL. In terms of acceleration, UniFL still exhibits notable performance advantages, surpassing the LCM with the same 4-step inference on both SD1.5 and SDXL. Surprisingly, we found that UniFL sometimes obtained even better aesthetic quality with fewer inference steps. For example, when applied to SD1.5, the aesthetic score is boosted from 5.26 to 5.54 without acceleration. After executing the acceleration with the adversarial feedback learning, the aesthetic score is further improved to 5.88 with much fewer inference steps. The related reasons will be investigated in the ablation experiment. We also compared the two latest acceleration methods on SDXL, including the SDXL Turbo and SDXL Lightning. Although retaining the high text-to-image alignment, we found that the image
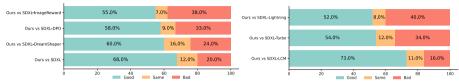
Fig. 4: The user study about UniFL and other methods with 10 users on the generation of 500 prompts in generation quality(right) and inference acceleration(right).

Fig. 5: The visualization of the generation results of different methods based on SDXL.

generated by SDXL Turbo tends to lack fidelity, leading to an inferior FID score. SDXL Lightning achieves the most balanced performance in all of these aspects and reaches impressive aesthetic quality in 4-step inference. However, UniFL still outperforms it in all kinds of metrics and achieves the best performance.

User Study. We conducted a comprehensive user study using SDXL to evaluate the effectiveness of our method in enhancing generation quality and acceleration. As illustrated in Fig.4, our method significantly improves the original SDXL in terms of generation quality with 68% preference rate and outperforms DreamShaper and DPO by 36% and 25% preference rate, respectively. Thanks to perceptual feedback learning and decoupled aesthetic feedback learning, our method exhibits improvement even when compared to the competitive ImageReward, and is preferred by 17% additional people. In terms of acceleration, our method surpasses the widely used LCM by a substantial margin of 57% with 4-step inference. Even when compared to the latest acceleration methods like SDXL-Turbo and SDXL-Lightning, UniFL still demonstrates superiority and obtains more preference. This highlights the effectiveness of adversarial feedback learning in achieving acceleration.

Qualitative Comparison As shown in Fig.5, UniFL achieves superior generation results compared with other methods. For example, when compared to ImageReward, UniFL generates images that exhibit a more coherent object struc-

Fig. 6: Illustration of PeFL's impact on structure optimization. The instance segmentation model(SOLO) provides more effective feedback on structure generation issues such as distorted limbs, missing key objects, and subject ambiguity during fine-tuning.

Fig. 7: The ablation study on the PeFL optimization. We compared the generated results utilizing diffusion loss and PeFL with the same perceptual dataset. (a) The effect of PeFL on style optimization. (b) The effect of PeFL on the structure optimization.

ture (e.g., the horse), a more appropriate style (e.g., the cocktail), and a more captivating aesthetic quality (e.g., the warrior). Notably, even with fewer inference steps, UniFL consistently showcases higher generation quality, outperforming other methods. It is worth noting that SDXL-Turbo, due to its modification of the diffusion hypothesis, tends to produce images with a distinct style.

5.3   Ablation Study

How PeFL works.
   To gain a better understanding of how PeFL works, we take the example of structure optimization with PeFL and visualize the intermediate results. As shown in Fig.6, the instance segmentation model effectively captures the overall structure of the generated object and successfully identifies the structural defects, such as the distorted limbs of the little girl, the missing skateboard, and the surfboard, and the ambiguity elephant and horse. Instead of assigning equal importance to each pixel with naive diffusion loss, this type of feedback enables the diffusion model to focus more on the specific structured concepts. We showcase some generation results after optimization style and structure via PeFL in Fig.7. It is evident that the PeFL significantly boosts style generation(e.g. 'fres-
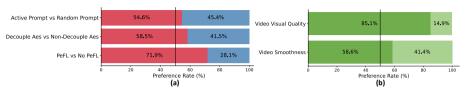
Fig. 8: (a) Human preference evaluation of different components in UniFL with SD1.5. (b) Human preference evaluation of the generated video via plugging the AnimiateDiff module into SD1.5(light green) and SD1.5 optimized by UniFL(dark green).

Fig. 9: Experiment of SD1.5 without and with the adversarial objective. (a) The intermediate results in 20 steps inference. (b) The 4 steps inference results.

cos', 'impasto' style) and object structure optimization(e.g. the woman's glasses, ballet dancer's legs) compared with the naive diffusion loss.

Effect of decoupled feedback learning.

To verify the importance of this decoupled aesthetic strategy, we conduct experiments by fine-tuning the SD1.5 model using a global aesthetic reward model trained with all the collected aesthetic preference data of different dimensions. As depicted in Fig.8(a), due to alleviate the challenge of abstract aesthetic learning, the utilization of decoupled aesthetic reward tuning resulted in generation results that were preferred by more individuals, surpassing the non-decoupled way by 17%. Fig.8(a) also shows that the active prompt selection obtained a higher preference rate(54.6% vs 45.4%), which demonstrates the importance of the prompt selection strategy.

How adversarial feedback learning accelerates. UniFL introduces adversarial feedback learning for acceleration, and the acceleration results even exceed the non-acceleration model in some scenarios; according to our experimental observations, the reason for the acceleration and the significant performance can be attributed to two potential factors: (i) Adversarial training enables the reward model to continuously provide guidance: As shown in Fig.9(a), the diffusion models with traditional feedback fine-tuning often suffer from rapid overfitting to the feedback signals generated by the frozen reward models, which is known as over-optimization. By employing adversarial feedback learning, the trainable reward model (acting as the discriminator) can swiftly adapt to the distribution shift of the diffusion model output, significantly mitigating the over-optimization phenomenon, which enables the reward model to provide effective guidance throughout the optimization process, (ii) Adversarial training expands

Fig. 10: Both SD1.5 and SDXL still keep high adaptation ability after being enhanced by the UniFL, even after being accelerated and inference with fewer denoising steps.

the time step of feedback learning optimization: Including the strong adversarial targets in the training process forces high-noise timesteps to generate clearer images via the adversarial objective, which enables the reward model to perform well even under a few denoising steps. As presented in Fig.9(b), after disabling the adversarial loss and retaining the optimization step containing the entire denoising process, the reward model fails to provide effective guidance for samples under fewer denoising steps without the adversarial training object due to the high-level noise, which results in poor inference results. With these two benefits, adversarial feedback learning significantly improves the generation quality of samples in lower inference steps and finally achieves superior acceleration performance.

For more ablation study of UniFL, please refer to the Appendix.

5.4   Generalization Study

To further verify the generalization of UniFL, we performed downstream tasks including LoRA, ControlNet, and AnimateDiff. Specifically, we selected several popular styles of LoRAs, several types of ControlNet, and AnimateDiff modules [18] and inserted them into our models respectively to perform corresponding tasks. As shown in Fig.10 and Fig.8(b), our model demonstrates excellent capabilities in style learning, controllable generation, and video generation.

6   Discussion and Limitations

UniFL demonstrates promising results in generating high-quality images. However, there are several avenues for further improvement:

Large Visual Perception Models: We are actively investigating the utilization of advanced large visual perception models to provide enhanced supervision.

Extreme Acceleration: While the current 1-step model's performance may be relatively subpar, the notable success we have achieved in 4-step inference suggests that UniFL holds significant potential for exploration in one-step inference.

Streamlining into a Single-stage Optimization: Exploring the possibility of simplifying our current two-stage optimization process into a more streamlined single-stage approach is a promising direction for further investigation.

7   Conclusion

We propose UniFL, a unified framework that enhances visual quality, aesthetic appeal, and inference efficiency through feedback learning. By incorporating perceptual, decoupled, and adversarial feedback learning, UniFL exceeds existing methods in terms of both generation quality and inference acceleration and generalizes well to various types of diffusion models and different downstream tasks.

References

1. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., Kaplan, J.: Training a helpful and harmless assistant with reinforcement learning from human feedback (2022) 4
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning (2024) 4
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions (2023) 4
4. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context (2018) 19
5. Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence (6), 679–698 (1986) 19
6. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023) 4
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017) 19, 20
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation (2022) 22, 23
9. civitai: Dreamshaper v8 (2024), https://civitai.com/models/4384/dreamshaper 9, 10

10. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., Tang, J.: Cogview: Mastering text-to-image generation via transformers (2021) 4

11. Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment (2023) 2

12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021) 10

13. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition (2021) 19

14. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity (2022) 2

15. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022) 2, 4

16. Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S.S., Shah, A., Yin, X., Parikh, D., Misra, I.: Emu video: Factorizing text-to-video generation by explicit image conditioning. arXiv preprint arXiv:2311.10709 (2023) 2

17. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014) 4

18. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023) 2, 14

19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022) 4

20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018) 10

21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021) 2, 4

22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022) 4

23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023) 19

24. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization (2019) 19

25. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation (2024) 9, 10

26. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015) 9

27. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference (2023) 2, 4, 9, 10

28. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module (2023) 2, 4, 10

29. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations (2022) 2, 4

30. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models (2022) 4

31. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022) 4

32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 2, 4, 9, 20

33. Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J.C., Xiong, C., Savarese, S., Ermon, S., Fu, Y., Xu, R.: Unicontrol: A unified diffusion model for controllable visual generation in the wild (2023) 2, 4

34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022) 2, 4

35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022) 2, 4

36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022) 9

37. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2023) 2, 4

38. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022) 2, 4

39. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models (2022) 2, 4

40. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation (2023) 2, 9, 10

41. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation (2023) 4

42. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023) 8

43. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022) 9

44. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer (2017) 2

45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015) 9

46. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023) 2, 3, 4

47. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., Dai, J., Qiao, Y., Wang, L., Li, H.: Journeydb: A benchmark for generative image understanding (2023) 20

48. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao,

C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023) 8, 24

49. Turc, I., Nemade, G.: Midjourney user prompts & generated images (250k) (2022). https://doi.org/10.34740/KAGGLE/DS/2349267 2, 20

50. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization (2023) 2, 9, 10

51. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Solo: A simple framework for instance segmentation (2021) 9, 23

52. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models (2023) 9

53. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation (2023) 2

54. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis (2023) 2, 4

55. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference (2023) 2, 4

56. Xie, S., Tu, Z.: Holistically-nested edge detection (2015) 19

57. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation (2023) 2, 3, 4, 5, 7, 9, 10

58. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. arXiv preprint arXiv:2311.09257 (2023) 8

59. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths (2023) 2, 4

60. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023) 2

61. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023) 2, 4

62. Zhang, Z., Zhang, S., Zhan, Y., Luo, Y., Wen, Y., Tao, D.: Confronting reward overoptimization for diffusion models: A perspective of inductive and primacy biases (2024) 8

63. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A., Chen, Z., Le, Q., Laudon, J.: Mixture-of-experts with expert choice routing (2022) 2

64. Štrupl, M., Faccio, F., Ashley, D.R., Srivastava, R.K., Schmidhuber, J.: Reward-weighted regression converges to a global optimum (2022) 4

UniFL: Improve Stable Diffusion via Unified Feedback Learning
(Appendix)

A   More Example of PeFL

The proposed perceptual feedback learning (PeFL) is very flexible and can leverage various existing visual perception models to provide specific aspects of visual quality feedback. To demonstrate the scalability of PeFL, we further provide a case study of PeFL for layout optimization based on the semantic segmentation model. Generally, the semantic segmentation map characterizes the overall layout of the image as shown in Fig.1(a). Therefore, semantic segmentation models can serve as a better layout feedback provider. Specifically, we utilize the visual semantic segmentation model to execute semantic segmentation on the denoised image x_0' to capture the current generated layout and supervise it with the ground truth segmentation mask and calculate semantic segmentation loss as the feedback on the layout generation:

$$\mathcal{L}(\theta)_{pefl\_layout} = \mathbb{E}_{x_0 \sim \mathcal{D}, x_0' \sim G(x_{t_a})} \mathcal{L}_{semantic}(m_s(x_0'), GT(x_0)) \tag{1}$$

where m_s represents the semantic segmentation model, $GT(x_0)$ is the ground truth semantic segmentation annotation and the $\mathcal{L}_{semantic}$ is the semantic segmentation loss depending on the specific semantic segmentation model.

We conduct the experiment of PeFL layout optimization based on SD1.5. Specifically, we utilize the COCO Stuff [4] with semantic segmentation annotation as the semantic layout dataset and DeepLab-V3 [7] as the semantic segmentation model. The experiment results are presented in Fix.1(b). It demonstrates that the PeFL significantly improves the layout of the generated image, for instance, the bear on the bed in a diagonal layout. We further conduct the user study to evaluate the effectiveness of PeFL with the semantic segmentation model quantitatively. The results are shown in Fig.4(c). Compared with the model fine-tuning with only the aesthetic feedback learning, incorporating the PeFL of layout optimization, the model generates images that obtain more preference in both layout and details terms.

Indeed, PeFL is an incredibly versatile framework that can be applied to a wide range of visual perceptual models, such as OCR models [13, 24] and edge detection models [5, 56], among others. Furthermore, we are actively delving into the utilization of the visual foundation model, such as SAM [23], which holds promising potential in addressing various visual limitations observed in current diffusion models.
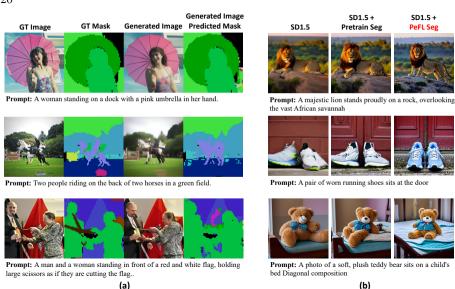
Fig. 1: (a) The illustration of the PeFL on the layout optimization. The semantic segmentation model captures the layout and text-to-image misalignment between the ground truth image and the generated image(DeepLab-V3 [7] is taken as the segmentation model). (b) The layout optimization effect of the PeFL with semantic segmentation model on SD1.5.

B Details of Data Collection in Decoupled Feedback Learning

We break down the general and coarse aesthetic concept into more specific dimensions including color, layout, detail, and lighting to ease the challenge of aesthetic fine-tuning. We then collect the human preference dataset along each dimension. Generally, we employ the SDXL [32] as the base model and utilize the prompts from the MidJourney [47, 49] as input, generating two images for each prompt. Subsequently, we enlist the expertise of 4 to 5 professional annotators to assess and determine the superior image among the generated pair. Given the inherently subjective nature of the judgment process, we have adopted a voting approach to ascertain the final preference results for each prompt. In addition to this manual annotation, we have combined two automatic feedback data generation methods to increase the diversity of preference data. Specifically, we initially curated a set of trigger words for each aesthetic dimension. By incorporating these words into input prompts, we can steer the diffusion model to focus on generating images that emphasize the corresponding aesthetic dimension. We then employ two strategies to generate the aesthetic feedback data, as depicted in Fig.3(b). (1) **Temporal Contrastive Samples**: As the denoising process progresses, it is typically observed that the model generates images of progressively higher quality. In the later denoising steps, the model demonstrates enhanced responsiveness to specific aesthetic trigger words. As a result, the image generated during the later denoising steps along with the one generated in the earlier denoising steps provides valuable feedback on the tar-

geted aesthetic dimension. (2) **Cross-Model Contrastive Samples**: Different diffusion models possess varying generative capabilities, resulting in images with different aesthetic qualities when provided with the same input prompt. We select several improved diffusion models, such as Kandinsky[*], and Playground[†], to generate preferred samples with higher aesthetic quality. Conversely, we use the vanilla diffusion model to generate unpreferred samples with inferior aesthetic quality. Combined with manual annotation and automatic annotation strategy, we finally curate 30,000, 32,000, 30,000, and 30,000 data points for the color, layout, detail, and lighting dimensions, respectively. Examples of the collected aesthetic feedback data of different dimensions are visually presented in Fig.2.

C   Extend Details of Prompt Selection

We introduce an active prompt selection strategy designed to choose the most informative and diverse prompts from a vast prompt database. The comprehensive implementation of this selection strategy is outlined in the Algorithm. Our strategy's primary objective is to select prompts that offer maximum information and diversity. To accomplish this, we have devised two key components: the **Semantic-based Prompt Filter** and the **Nearest Neighbor Prompt Compression**. The semantic-based prompt filter is designed to assess the semantic relationship embedded within the prompts and eliminate prompts that lack substantial information. To accomplish this, we utilize an existing scene graph parser[‡] as a tool to parse the grammatical components, such as the subjective and objective elements. The scene graph parser also generates various relationships associated with the subjective and objective, including attributes and actions. We then calculate the number of relationships for each subjective and objective and select the maximum number of relationships as the metric of information encoded in the prompt. A higher number of relationships indicates that the prompt contains more information. We filter out prompts that have fewer than $\tau_1 = 1$ relationships, which discard the meaningless prompt like 'ff 0 0 0 0' to reduce the noise of the prompt set.

Upon completing the filtration process, our objective is to select a predetermined number of prompts that exhibit maximum diversity. This approach is driven by the understanding that during preference fine-tuning, there is a tendency to encounter the over-optimization phenomenon, as depicted in Fig.4(c). In such cases, the diffusion model rapidly overfits the guidance provided by the reward model, ultimately resulting in the loss of effectiveness of the reward model. One contributing factor to this phenomenon is the distribution of prompts for optimization. If the prompts are too closely distributed, the reward model is forced to frequently provide reward signals on similar data points. This leads to the diffusion model rapidly overfitting and collapsing within a limited number of optimization steps. To address this challenge, we propose the selection of

---

[*] https://github.com/ai-forever/Kandinsky-2
[†] https://huggingface.co/playgroundai/playground-v2-1024px-aesthetic
[‡] https://github.com/vacancy/SceneGraphParser

Model	Style Response Rate
SD1.5	30.55%
SD1.5 + Style Pretrain	35.25%
SD1.5 + Style PeFL	**45.14%**
SDXL	66.67%
SDXL + Style Pretrain	68.34 %
SDXL + Style	**75.27%**

Table 1: Ablation on the PeFL in style generation.

Method	FID↓	CLIP Score↑	Aesthetic ↑
SD1.5	37.99	0.308	5.26
SD1.5 + Non-Decouple Aes	35.12	0.308	5.33
SD1.5 + Decouple Aes	33.78	0.310	5.40
SD1.5 + Decouple Aes + Active	**31.89**	**0.315**	**5.48**

Table 2: Ablation on the decoupled feedback learning. Decouple Aes: Decoupled Aesthetic Strategy; Active: Active Prompt Selection.

diverse prompts to mitigate the problem. Specifically, with a fixed number of prompt selections in mind, we aim to ensure that the chosen prompts exhibit maximum diversity. We adopt an iterative process to achieve this objective. In each iteration, we randomly select a seed prompt and subsequently suppress its nearest neighbor[§] prompts that have a similarity greater than $\tau_2 = 0.8$. The next iteration commences with the remaining prompts, and we repeat this process until the similarity between the nearest neighbors of all prompts falls below the threshold τ_2. Finally, we randomly select the prompts, adhering to the fixed number required for preference fine-tuning.

D   More Ablation Study

Extend ablation on the PeFL. We conduct a further detail ablation study to evaluate the effectiveness of the PeFL on style optimization. Specifically, we collect 90 prompts about style generation and generate 8 images for each prompt. Then, we manually statistic the rate of correctly responded generation to calculate the style response rate. As presented in Tab.1, it is evident that the style PeFL greatly boosts the style generation on both architectures, especially for SD1.5 with about 15% improvement. In contrast, leveraging naive diffusion loss for fine-tuning with the same collected style dataset suffers limited improvement due to stylistic abstraction missing in latent space.

Ablation on Visual Perceptual Model Selection. PeFL utilizes various visual perceptual models to provide visual feedback in specific dimensions to improve the visual generation quality on particular aspects. Different visual perceptual models of a certain dimension may have different impacts on the performance of PeFL. Taking the structure optimization of PeFL as an example, we investigated the impact of the accuracy of instance segmentation models on PeFL performance. Naturally, the higher the precision of the instance segmentation, the better the performance of structure optimization. To this end, we choose the Mask2Former [8], another representative instance segmentation model with state-of-the-art performance, for the structure optimization with PeFL. The results are shown in Fig.3(a) and Fig.4(b). It is intriguing to note that the utilization of a higher precision instance segmentation model does not

---

[§] https://github.com/facebookresearch/faiss

Algorithm 2 Active Prompt Selection

1: **Input:** Initial collected prompt set $\mathcal{D}$.
2: **Initialization:** The number of selected prompts N for aesthetic preference fine-tune, decided by the optimization steps until overoptimization.
3: **Return:** The final selected prompts SP
4: $\mathcal{P} = \varnothing$ // Initialize the filtered prompts set
5: **for** prompt $p_i \in \mathcal{D}$ **do**
6:      $SR \leftarrow \text{SemanticParser}(p_i)$
7:      **if** $|SR| > \tau_1$ **then**
8:          $\mathcal{P} \leftarrow p_i$ // Append the informative prompt
9:      **end if**
10: **end for**
11: $I \leftarrow \text{shuffle}(\text{range}(\text{len}(|\mathcal{P}|)))$ // Get the random index
12: $R \leftarrow \text{False}$ // Set the removed prompt array
13: $S \leftarrow \varnothing$ // Set the selected prompt index
14: Dist, Inds $\leftarrow \text{KNN}(R, k)$ // Get the K-nearest neighbor for each prompt
15: **for** index $I_i \in I$ **do**
16:      **if** not $R[I_i]$ and I_i not in S **then**
17:          $S \leftarrow I_i$ // Append the selected prompt
18:          dist, inds $= \text{Dists}[I_i], \text{Inds}[I_i]$ // Get the K-nearest neighbor similarity
19:          **for** index $d_i \in$ inds **do**
20:              **if** $\text{dist}[d_i] > \tau_2$ **then**
21:                  $R[d_i] = \text{True}$
22:              **end if**
23:          **end for**
24:      **end if**
25: **end for**
26: $SP \leftarrow \text{RandomSelect}(\mathcal{P}, S, N)$ // Random select N diverse prompt according the retained index
27: **return** SP

---

yield significantly improved results in terms of performance. This may be attributed to the different architectures of the instance segmentation of these two models. In SOLO [51], the instance segmentation is formulated as a pixel-wise classification, where each pixel will be responsible for a particular instance or the background. Such dense supervision fashion enables the feedback signal to better cover the whole image during generation. In contrast, Mask2Former [8] takes the query-based instance segmentation paradigm, where only a sparse query is used to aggregate the instance-related feature and execute segmentation. This sparse nature of the query-based method makes the feedback insufficient and leads to inferior fine-tuning results. We leave further exploration of how to choose the most appropriate visual perceptual model for feedback tuning to future work.

Ablation on Decoupled Feedback Learning. To evaluate the effectiveness of the two key designs in decoupled feedback learning, namely decoupled aesthetic feedback fine-tuning and active prompt selection, we conducted a series of incremental experiments using SD1.5. The results of these experiments are summarized in Tab.2. To validate the necessity of the decoupled design, we first

trained a global aesthetic reward model using the collected aesthetic preference data across different dimensions and directly fine-tuned the diffusion model using this reward model. As shown in Tab.2, while direct fine-tuning yielded some reasonable improvements, the decoupled aesthetic reward models achieved more significant performance enhancements, particularly in terms of FID and aesthetic quality (FID: 33.78 vs 35.12, aesthetic: 5.40 vs 5.26). This is because decoupled design not only reduces the complexity of learning abstract aesthetic concepts but also mitigates potential conflicts during optimization as studied in [48]. Building upon the decoupled reward model fine-tuning, the incorporation of active prompt selection resulted in a further boost in performance. This emphasizes the crucial role of prompt selection in aesthetic fine-tuning and demonstrates its importance in achieving superior results. Actually, as depicted in Fig.4(a), we observed that the over-optimization can be greatly eased with the actively selected prompts. Such informative and diverse prompts allow the reward model to provide feedback under a broader data distribution, avoiding overfitting quickly and finally optimizing the diffusion model more sufficiently.

Ablation on Acceleration Steps. We comprehensively compared UniFL and existing acceleration methods using different numbers of inference steps, ranging from 1 to 8 as illustrated in Fig.5. Generally, UniFL performs exceptionally well with 2 to 8 inference steps, achieving superior text-to-image alignment and higher aesthetic quality. The LCM method is prone to generate blurred images when using fewer inference steps and requires more steps (e.g., 8 steps) to produce images of reasonable quality. However, both UniFL and LCM struggle to generate high-fidelity images with just 1-step inference, exhibiting a noticeable gap compared to SDXL-Turbo. This discrepancy arises because SDXL-Turbo is intentionally designed and optimized for extremely low-step inference scenarios. Consequently, when more inference steps are employed (e.g., in the case of the Labradoodle), the color of the image tends to appear unnatural. Therefore, there is still room for further exploration to enhance the acceleration capabilities of UniFL towards 1-step inference.

E   More Visualization Results

We present more visual comparison between different methods in Fig.6. It demonstrates the superiority of UniFL in both the generation quality and the acceleration. In terms of generation quality, UniFL exhibits more details(e.g. the hands of the chimpanzee), more complete structure(e.g. the dragon), and more aesthetic generation(e.g. the baby sloth and the giraffe) compared with DPO and ImageReward. In terms of acceleration, the LCM tends to generate a blurred image, while the SDXL-Turbo generates the image with an unpreferred style and layout. As a comparison, UniFL still retains the high aesthetic detail and structure under the 4-step inference.

Preferred   **Unpreferred**
A female sleuth with a hat, ultrhigh resolution

Preferred   **Unpreferred**
Galadriel crossing Khazadum. Lord of the rings style

Color Feedback Data

Preferred   **Unpreferred**
royal avenue belfast, illustrated art, white background, realistic, 3D

Preferred   **Unpreferred**
Lerson pannawit style, hologram holographic vibrant bizzare odd creepy weirdcore, photography bizzare, 90s magazine collage surrealism

Layout Feedback Data

Preferred   **Unpreferred**
Roses of black color surround a black-painted goat skull, black and white coloring page, scribbly, messy, 2D poster

Preferred   **Unpreferred**
Wind turbine style painting PS Krøyer

Detail Feedback Data

Preferred   **Unpreferred**
Whispering mechanical tree

Preferred   **Unpreferred**
3 Warm yellow, lightful, minimalistic, flying floating gift boxes, warm light yellow gradient background, frontal view

Lighting Feedback Data

Fig. 2: The visualization of the collected aesthetic feedback data along different dimensions.
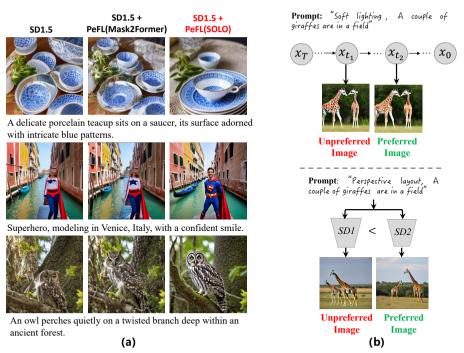
Fig. 3: (a) The visual comparison between the PeFL structure optimization with different instance segmentation models. (b) The illustration of the automatic aesthetic feedback data annotation. Top: Temporal Contrastive Samples. Bottom: Cross-Model Contrastive Samples.
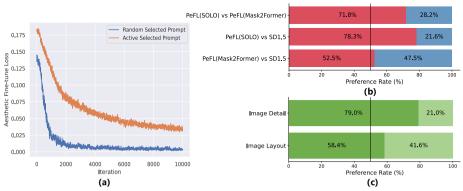

Fig. 4: (a) The visualization of the training loss curve during the preference fine-tuning of different prompt sets. (b) The user study results on the PeFL instruction optimization with different instance models. **PeFL(SOLO)**: PeFL fine-tune SD1.5 with SOLO as instance model. **PeFL(Mask2Former)**: PeFL fine-tune SD1.5 with Mask2Former as instance model. (c) The user study results on the SD1.5 with (Dark Green) and without(Light Green) PeFL layout optimization.
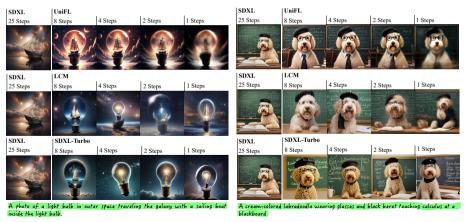
Fig. 5: The visual comparison of different acceleration methods under various inference steps.

Fig. 6: More visualization of the generated results of different methods.