UniFL: Improve Latent Diffusion Model via Unified Feedback Learning

Jiacheng Zhang, Jie Wu, Yuxi Ren, Xin Xia, Huafeng Kuang, Pan Xie, Jiashi Li,

Xuefeng Xiao, Weilin Huang, Shilei Wen, Lean Fu, Guanbin Li

Motivation

Despite the remarkable advancement in the diffusion-based text-to-image (T2I) generation, several limitations still exist in current latent diffusion models:

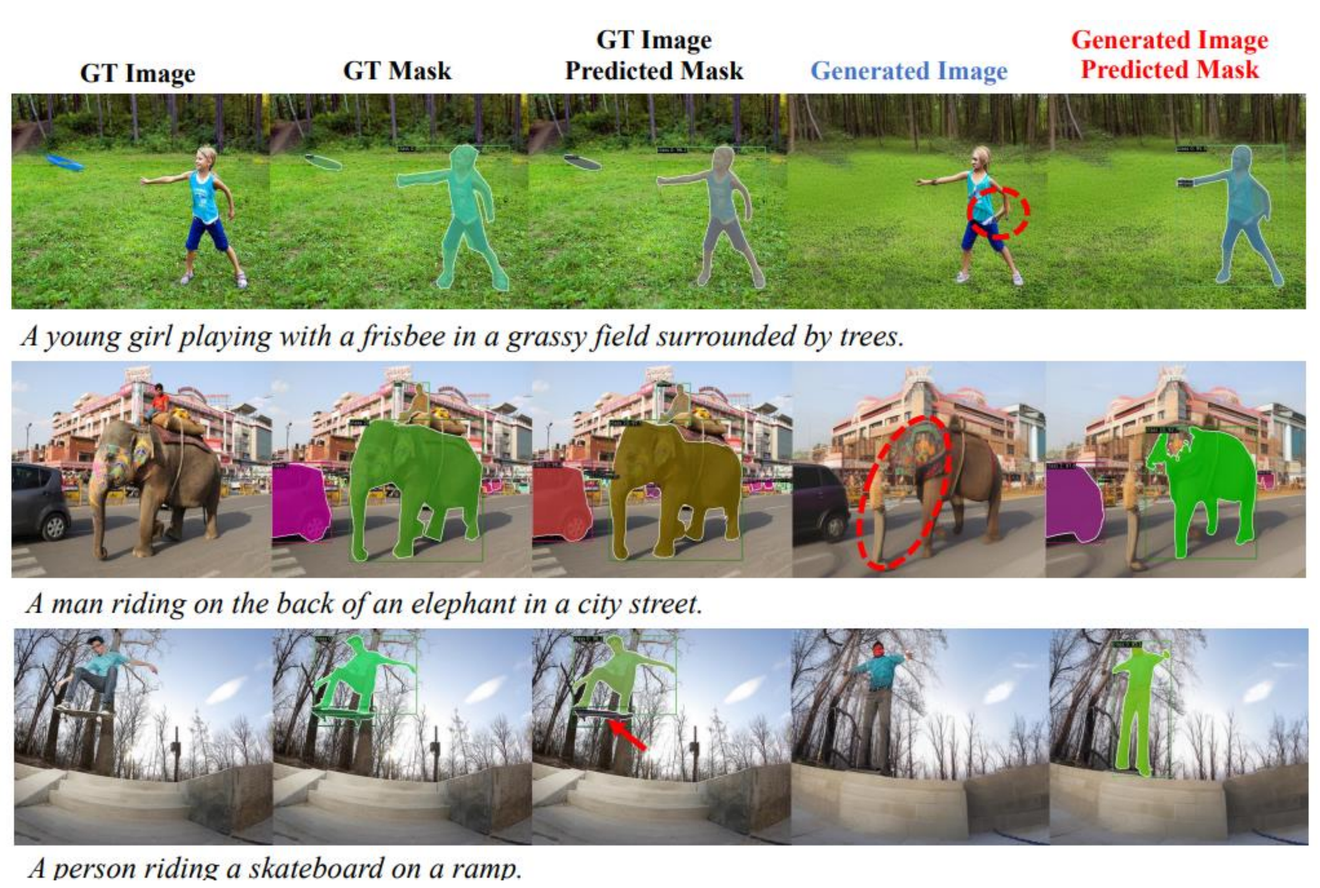
- *Inferior visual quality*: Poor visual quality and lack authenticity, e.g. characters with incomplete limbs or distorted body parts.
- *Inadequate aesthetic appeal*: The generated image tends to lack aesthetic appeal and often fails to align with human preferences,
- *Slow inference speed:* Considerable iterative denoising steps are required to obtain the decent generation results.

There are already some methods that concentrate on tackling individual problems through specialized design, but no method to tackle these problems comprehensively with a unified design.

UniFL: Unified Feedback Learning for LDMs

UniFL aims to improve the latent diffusion models in various aspects, including visual generation quality, human aesthetic quality, and inference efficiency from the unified perspective of feedback learning.

Perceptual Feedback Learning: Repurpose the pretrained perceptual models to provide more specifical and targeted feedback supervision.



For example, the instance segmentation model (here is SOLO[1]) can assist in providing the feedback supervision on the visual structure completeness

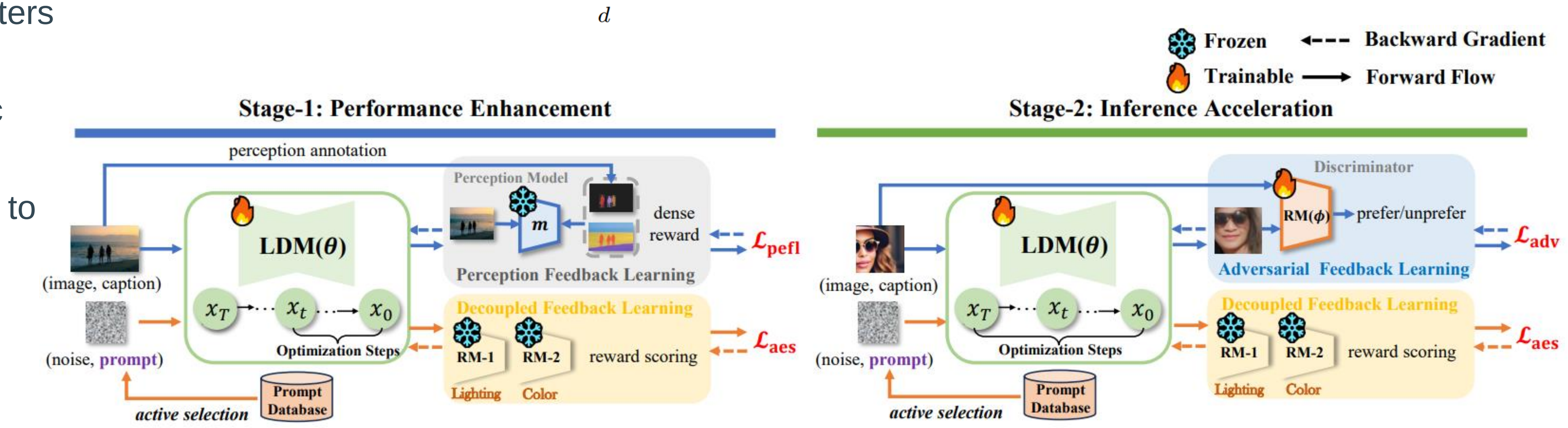
Project Page

Arxiv

[1] SOLO: Segmenting Objects by Locations. European Conference on Computer Vision 2020

Decoupled Feedback Learning: Human preference feedback is a multi-dimensional nature, which requires separate reward models to model the different aspects of human aesthetic preference and impose preference fine-tuning. We curate the human aesthetic preference model on: color, layout, detail, lighting

$$\mathcal{L}_{\text{aes}}(\theta) = \sum_{d}^{K} \mathbb{E}_{c\sim p(c)}\mathbb{E}_{x'_0\sim p(x'_0|c)}[\texttt{ReLU}(\alpha_d - r_d(x'_0, c))]$$



Overview of the UniFL framework. It is instantiated by a two-stage training process, with the first stage focusing on the overall quality (objective visual quality and subjective aesthetic quality) enhancement and the later stage for the inference acceleration.

Adversarial Feedback Learning: Incorporate the adversarial objective with the reward feedback learning, and enable the optimization for the images that go through lower denoising steps, leading to reasonable generation performance with fewer denoising steps and achieving inference acceleration.

$$\mathcal{L}^G(\theta) = \mathbb{E}_{c\sim p(c)}\mathbb{E}_{x'_0\sim p(x'_0|c)}[-r_a(x'_0, c)],$$

$$\mathcal{L}^D(\phi) = -\mathbb{E}_{(x_0,x'_0,c)\sim\mathcal{D}_{\mathbf{train}},t\sim[1,T]}[\log\sigma(r_a(x_0)) + \log(1-\sigma(r_a(x'_0)))].$$

Training Objective: The complete training objective with UniFL is:

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

Experiments

Experiment with SD1.5 and SDXL, and compare with the method that focuses on the generation performance enhancement and the method that focuses on the inference acceleration.

Algorithm 1 Perceptual Feedback Learning (PeFL)

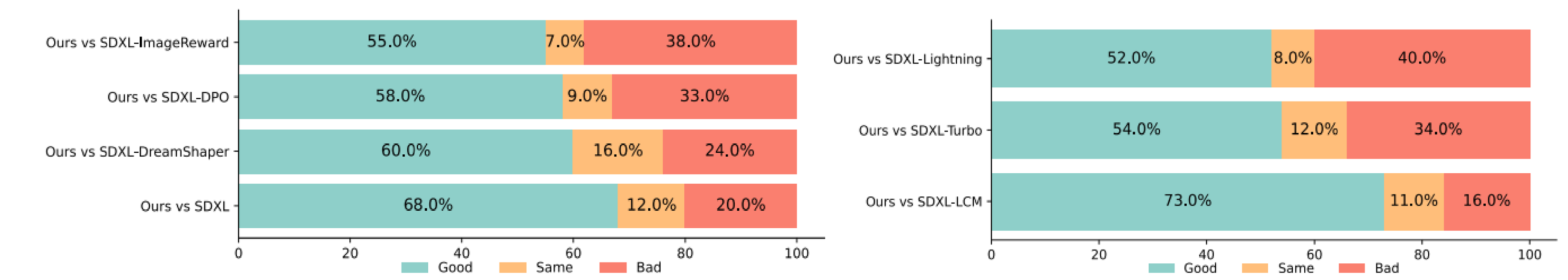
1: **Dataset:** Captioned perceptual text-image dataset with $\mathcal{D} = \{(\text{txt}_1, \text{img}_1), ...(\text{txt}_n, \text{img}_n)\}$
2: **Input:** LDM with pre-trained parameters w_0, perceptual model $m_\cdot$, perceptual loss function Φ, loss weight λ
3: **Initialization:** The number of 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: $x_0 \leftarrow \text{VaeEnc}(\text{img}_i)$ // From image to latent
6: $x_{T_a} \leftarrow$ **AddNoise**(x_0) // Add noise to latent
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 denoised latent
12: $\text{img}'_i \leftarrow \text{VaeDec}(x'_0)$ // From latent to image
13: $\mathcal{L}_{\text{pefl}} \leftarrow \lambda\Phi(m(\textbf{img}'_i), \text{GT}(\textbf{img}_i))$ // PeFL loss by perceptual model
14: $w_{i+1} \leftarrow w_i$ // Update LDM_{w_i} using PeFL loss
15: **end for**

The complete procedure of our perceptual feedback learning (PeFL)

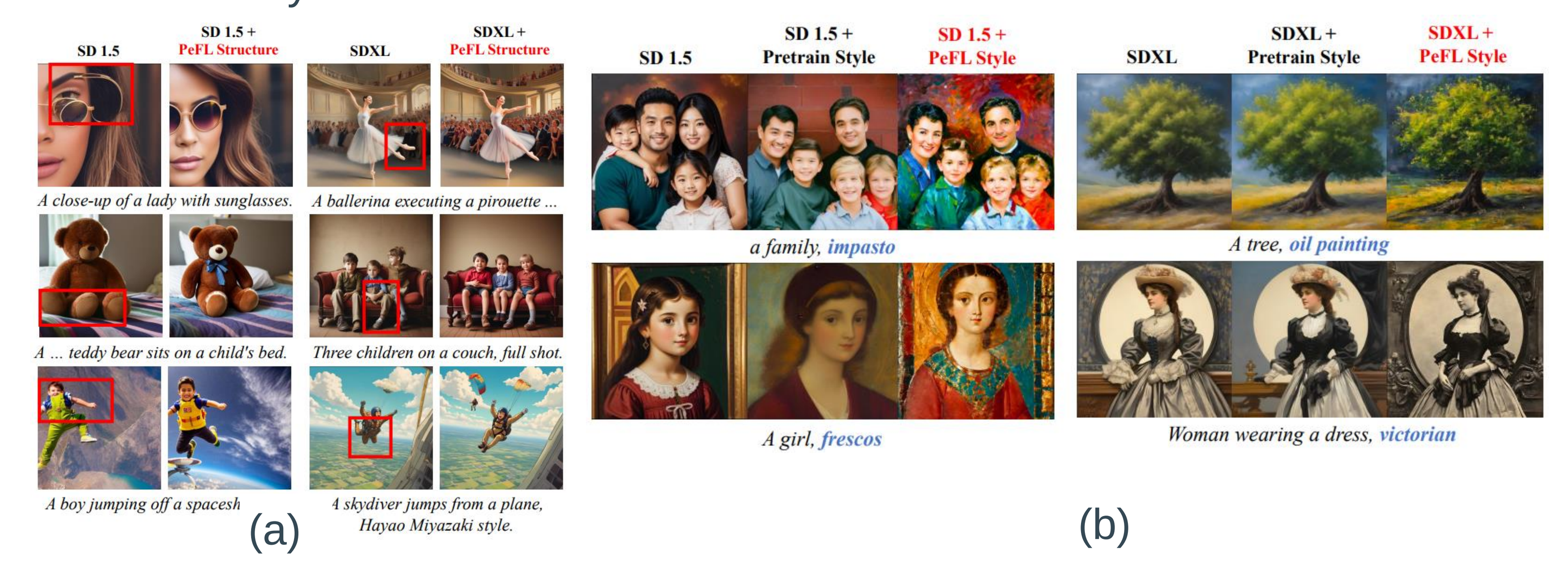
Model	Step	FID↓	CLIP Score↑	Aes Score↑
SD15-Base	20	37.99	0.308	5.26
SD15-IR [23]	20	32.31	0.312	5.37
SD15-DS [52]	20	34.21	0.313	5.44
SD15-DPO [22]	20	32.83	0.308	5.22
SD15-UniFL	20	**31.14**	**0.318**	**5.54**
SD15-Base	4	42.91	0.279	5.16
SD15-LCM [27]	4	42.65	0.314	5.71
SD15-DS LCM [26]	4	35.48	0.314	5.58
SD15-UniFL	4	**33.54**	**0.316**	**5.88**
SDXL-Base	25	27.92	0.321	5.65
SDXL-IR [23]	25	26.71	0.319	5.81
SDXL-DS [52]	25	28.53	0.321	5.65
SDXL-DPO [22]	25	35.30	0.325	5.64
SDXL-UniFL	25	**25.54**	**0.328**	**5.98**
SDXL-Base	4	125.89	0.256	5.18
SDXL-LCM [27]	4	27.23	0.322	5.48
SDXL-Turbo [24]	4	30.43	0.325	5.60
SDXL-Lighting [53]	4	28.48	0.323	5.66
SDXL-UniFL	4	**26.25**	**0.325**	**5.87**

The quantitative performance comparison with other methods.

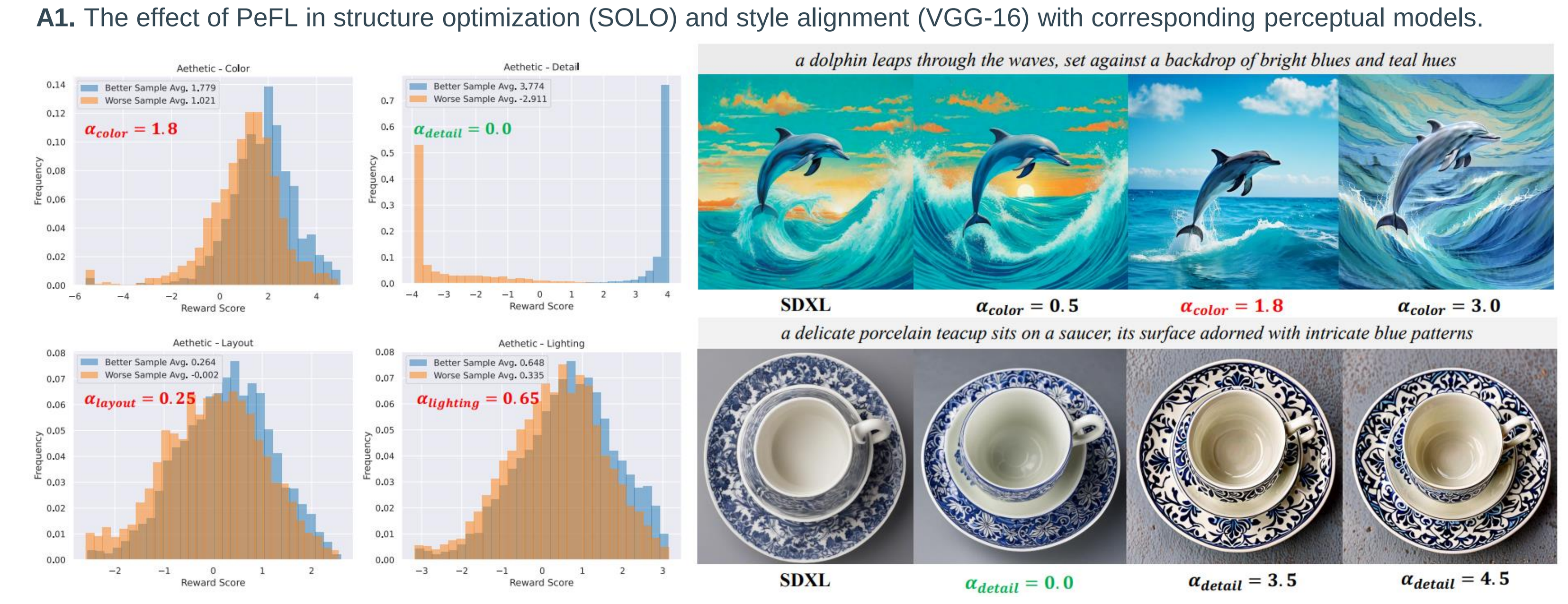
User study on both generation quality enhancement and inference acceleration.



Ablation study

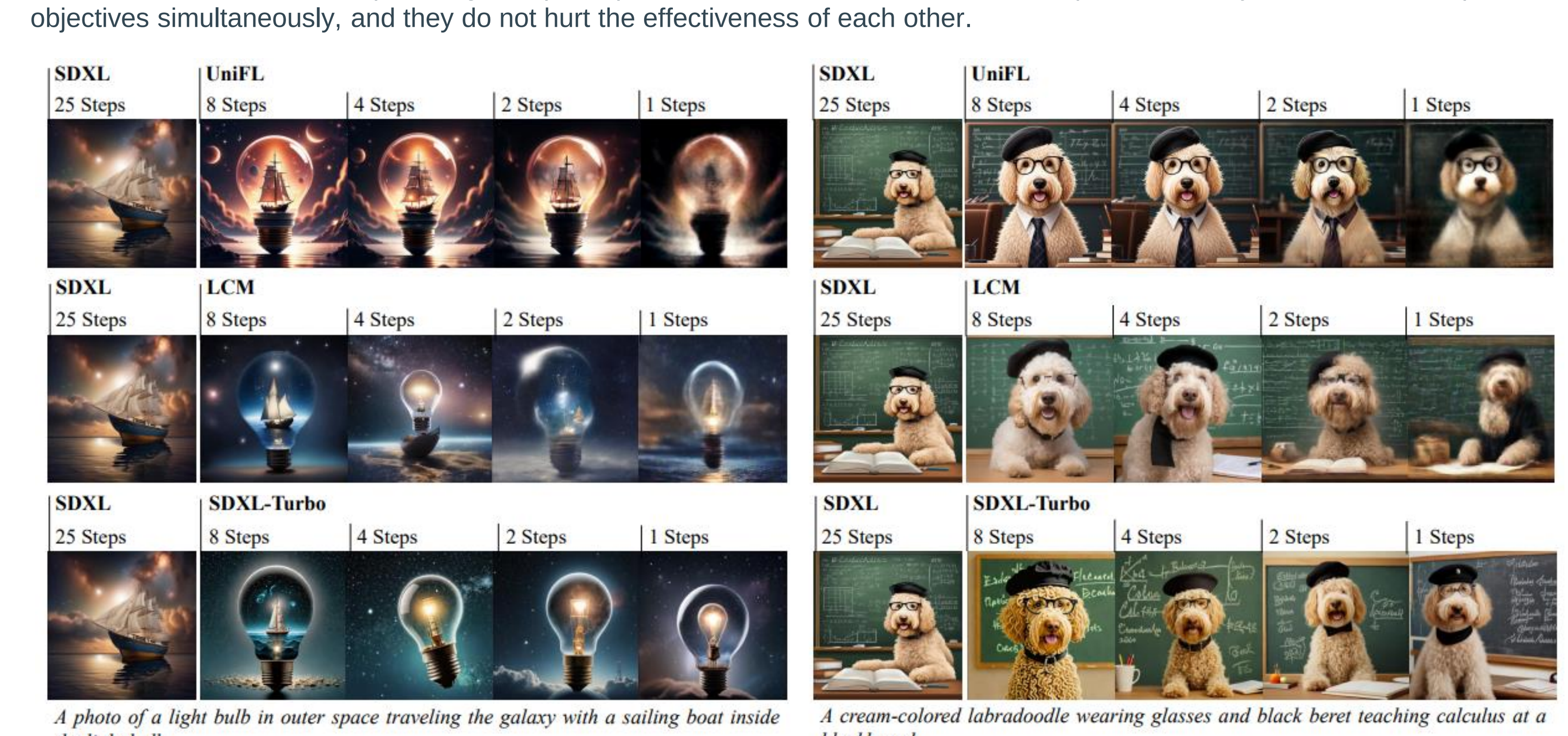


A1. The effect of PeFL in structure optimization (SOLO) and style alignment (VGG-16) with corresponding perceptual models.



A2. The selection of the hinge coefficients α_d for different aesthetic dimensions. The discrepancy also highlights the necessity of decoupled design.

A3. PeFL also succeeds in optimizing multiple aspects at the same time. Here, we incorporated the style and structure optimization objectives simultaneously, and they do not hurt the effectiveness of each other.



A4. UniFL is superior in inference acceleration with 2-8 steps, and still inferior in single steps.