Self-Rewarding Large Vision-Language Models for Optimizing Prompts in Text-to-Image Generation

Hongji Yang, Yucheng Zhou, Wencheng Han, Jianbing Shen[*]

SKL-IOTSC, CIS, University of Macau

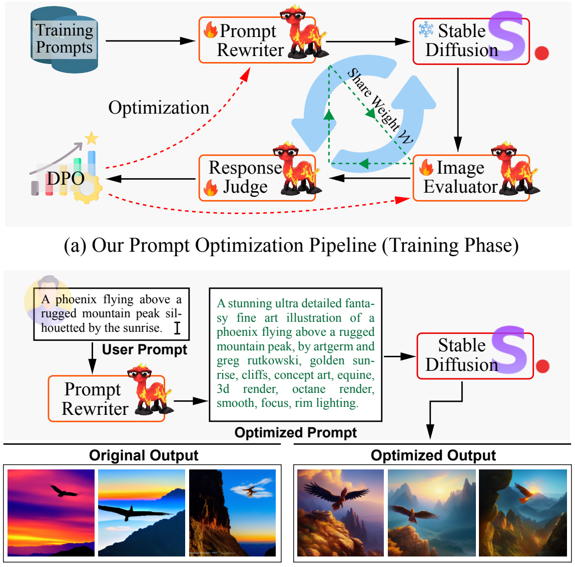
yc47942@um.edu.mo, yucheng.zhou@connect.um.edu.mo, wencheng256@gmail.com

Abstract

Text-to-image models are powerful for producing high-quality images based on given text prompts, but crafting these prompts often requires specialized vocabulary. To address this, existing methods train rewriting models with supervision from large amounts of manually annotated data and trained aesthetic assessment models. To alleviate the dependence on data scale for model training and the biases introduced by trained models, we propose a novel prompt optimization framework, designed to rephrase a simple user prompt into a sophisticated prompt to a text-to-image model. Specifically, we employ the large vision language models (LVLMs) as the solver to rewrite the user prompt, and concurrently, employ LVLMs as a reward model to score the aesthetics and alignment of the images generated by the optimized prompt. Instead of laborious human feedback, we exploit the prior knowledge of the LVLM to provide rewards, i.e., AI feedback. Simultaneously, the solver and the reward model are unified into one model and iterated in reinforcement learning to achieve self-improvement by giving a solution and judging itself. Results on two popular datasets demonstrate that our method outperforms other strong competitors.

1 Introduction

Text-to-image models (Rombach et al., 2022; Saharia et al., 2022; Yu et al., 2022) can generate diverse high-quality images based on user-provided prompts. However, effective interaction with these models requires users to possess specific expertise, including familiarity with specialized vocabularies, e.g., "*35mm*" for camera parameters and "*art by Greg Rutkowski*" to invoke a particular artistic style. As shown in Figure 1, rewriting prompts according to model-specific knowledge significantly improves the quality of the generated images.



(a) Our Prompt Optimization Pipeline (Training Phase)

A phoenix flying above a rugged mountain peak silhouetted by the sunrise.

A stunning ultra detailed fantasy fine art illustration of a phoenix flying above a rugged mountain peak, by artgerm and greg rutkowski, golden sunrise, cliffs, concept art, equine, 3d render, octane render, smooth, focus, rim lighting.

(b) Our Prompt Optimization Pipeline (Inference Phase)

Figure 1: The motivation of our prompt optimization pipeline. When training, the large model continuously improves prompt rewriting and image quality evaluation capabilities through self-play without any external sources or models. The generated images from the user prompts and the rewritten prompts. It can be observed that the image from the modified prompt has higher aesthetics.

To bridge the gap between laymen and experts in using text-to-image models, some methods (Liu and Chilton, 2022; Datta et al., 2023; Oppenlaender, 2023) employ large-scale human-annotated datasets to train rewriting models that produce more professional and effective prompts. However, these approaches are prohibitively expensive. To reduce reliance on high-quality human annotations, some efforts (Hao et al., 2024; Rosenman et al., 2023; Cao et al., 2023) leverage specific calculated metrics, i.e., aesthetics and alignment, which are then regarded as rewarding for reinforcement learning (RL). However, collecting high-quality human-annotated datasets for reward model training is

[*] Corresponding author. This work was supported by the Science and Technology Development Fund of Macau SAR (FDCT) under grants 0102/2023/RIA2 and 0154/2022/A3 and 001/2024/SKL and CG2025-IOTSC, the National Natural Science Foundation of China (No. 624B2002), and the Jiangyin Hi-tech Industrial DevelopmentZone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC).

Findings of the Association for Computational Linguistics: ACL 2025, pages 7332–7349
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

both time-consuming and costly. Large Vision-Language Models (LVLMs) have recently shown remarkable vision reasoning capabilities. Consequently, some studies (Li et al., 2022; Liu et al., 2024; Chen et al., 2024a; Yang et al., 2023; Yu et al., 2023) employ LVLMs as evaluators to assess human preferences. These methods effectively deliver interpretable AI feedback, offering a more efficient alternative to the time-consuming and labor-intensive human feedback (Yuan et al., 2024a).

Besides, previous prompt optimization methods using RL suffer from two limitations: (1) They require extensive training data to train an image reward model; (2) The reward model remains fixed during the proxy model training, preventing it from learning and improving alongside the proxy model. This limitation results in a lack of dynamic feedback throughout the training process. To mitigate the data limitation and explore dynamic feedback, Yuan et al. (2024b) introduce a self-rewarding training strategy, enabling the model to train effectively with limited data while approximating the upper performance bound. It fosters self-improvement or self-play (Chen et al., 2024b), wherein the solver generates its judgments or rewards by a continuous iterative DPO (Xu et al., 2023).

In this study, we introduce a self-rewarding prompt optimization framework for text-to-image models. This framework leverages an LVLM which functions both as a solver and an evaluator. The training pipeline is structured into five key components: model initialization, prompt generation, image generation, LVLM rewarding, and RL training. The pipeline is as follows: (1) Model initialization: we train the LVLM for prompt optimization on human-annotated prompt rewriting pairs and limited evaluation data, encapsulating LVLM's basic capability to rewrite prompts and assess the preference of generated images. (2) Prompt Generation: we employ a large version of LVLM to generate responses for a combination of an existing image quality evaluation dataset with the image evaluation prompt we used. (3) Image Generation: the model rewrites the raw prompt according to the instruction and samples multiple results, i.e., "candidates", and a fixed text-to-image model is used to generate the corresponding image. (4) LVLM Rewarding: the LVLM is employed to evaluate the aesthetics and alignment of the image with the original prompt, and is utilized as a rating system to score the images due to endowing with the capability to assess preferences. (5) RL Training: we select the highest and lowest-scored candidates to form preference pairs, which are used to train the model and adjust its output preferences by DPO training. To further enhance the model's ability to judge image quality, we also make the model itself act as a judge on the model's responses to image aesthetics or alignment to pick the most and least confident responses to construct preference pairs.

The main contributions are summarized below:

- We provide an AI-feedback approach to achieve aesthetic and alignment understanding of images using an LVLM. This approach effectively transforms the LVLM into a reward model to facilitate prompt optimization.
- We introduce self-rewarding training into prompt optimizing for the first time, which obtained prompts with higher quality by iterating the model on a small amount of training data, alleviating the shortcomings of models that require larger and higher quality data for reinforcement learning training.
- In the experiments, we compare our method and other strong competitors on two popular text-to-image datasets, i.e., beautiful-prompt and DiffusionDB. Results show our method achieves state-of-the-art performance.

2 Background

2.1 Prompt Rewriting

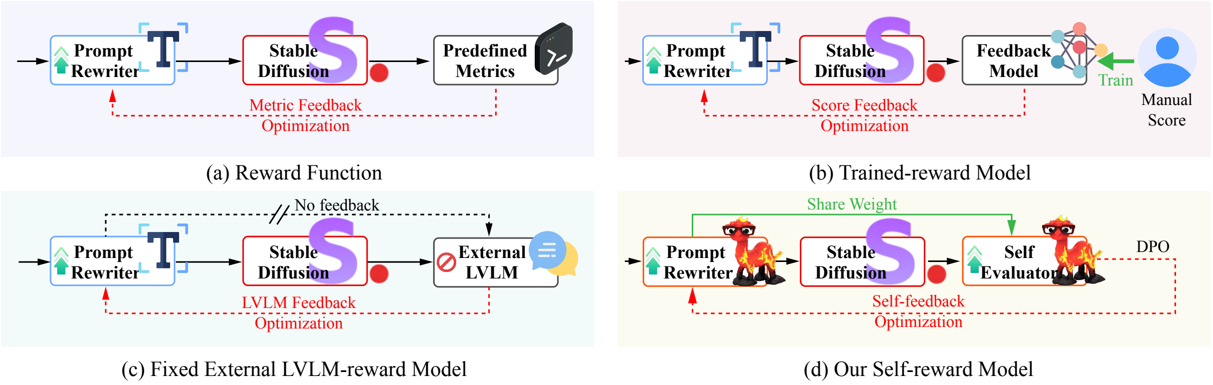
The purpose of the prompt rewriting is to unlock the maximum potential of text-to-image models. Given an original user prompt $\mathbf{x}$, the prompt rewrite models with weight θ can produce more professional prompts $\mathbf{y}$ to help text-to-image models achieve images with more aesthetic pleasure and relevance. The process can be expressed as follows:

$$\mathbf{y} = p(y|\mathbf{x};\theta) \tag{1}$$

2.2 Self-Rewarding Learning

Self-Rewarding Learning uses the same model to perform iterative training to realize self-improvement. Given a sequence pair (x, y), reinforcement learning fine-tuning demands the definition of the reward function $\mathbf{r}(\mathbf{x}, \mathbf{y})$ to quantify the value of the response $\mathbf{y}$ to the given input $\mathbf{x}$. The objective of the RL fine-tuning can be defined as:

$$\begin{aligned}\mathcal{L}_{RL}(\theta) = &\ \mathbb{E}_{\mathbf{x}\sim q(\cdot),\mathbf{y}\sim p_\theta(\cdot|\mathbf{x})}[r(\mathbf{x},\mathbf{y})] \\ &- \lambda\mathbb{E}_{\mathbf{x}\sim q(\cdot)}\mathrm{KL}(p_\theta(\cdot|\mathbf{x})||p_{ref}(\cdot|\mathbf{x}))\end{aligned} \tag{2}$$



(a) Reward Function

(b) Trained-reward Model

(c) Fixed External LVLM-reward Model

(d) Our Self-reward Model

Figure 2: **Four types of framework in prompt optimizing**. The main difference among them is the reward generation. **(a)** The reward function (or Metrics) is pre-defined for reinforcement learning, which typically involves mathematical equations for text or images, such as work (Hao et al., 2024). **(b)** A feedback model is trained using a large annotated dataset (typical manual score) and then is employed for RL, like PPO, which is employed in (Cao et al., 2023; Rosenman et al., 2023). **(c)** Rewards are generated through a **fixed external LVLM**. The upper bound of evaluation depends on external models. **(d)** Rewriter and Reward Model share the same weight, achieving self-improvement by an iterative method to generating answers and self-judgment.

where the KL regularization enforces the policy model p_θ to be close to the reference model p_{ref}, and λ is set to control the deviation between policy model p_θ and the reference model p_{ref}.

Assuming a pair of responses $< \mathbf{y}^1, \mathbf{y}^2 >$, with a human annotator labeling one of them to be more aligned with human preferences, denoted as $\mathbf{y}^w \succ \mathbf{y}^l | x$. the Bradley-Terry (BT) (Brown et al., 2020) model stipulates that the human preference distribution p^* can be written as follows:

$$p^*(\mathbf{y}^1 \succ \mathbf{y}^2 | \mathbf{x}) = \frac{\exp(r(\mathbf{x}, \mathbf{y}^1))}{\exp(r(\mathbf{x}, \mathbf{y}^1)) + \exp(r(\mathbf{x}, \mathbf{y}^2))} \quad (3)$$

In self-rewarding training, the same LVLM is used to produce reward $\mathbf{r} = [r_1, r_2, ..., r_l]$, so the conditional probability distribution $p_\theta(r|(\mathbf{x}, \mathbf{y}))$ can be expressed as follows:

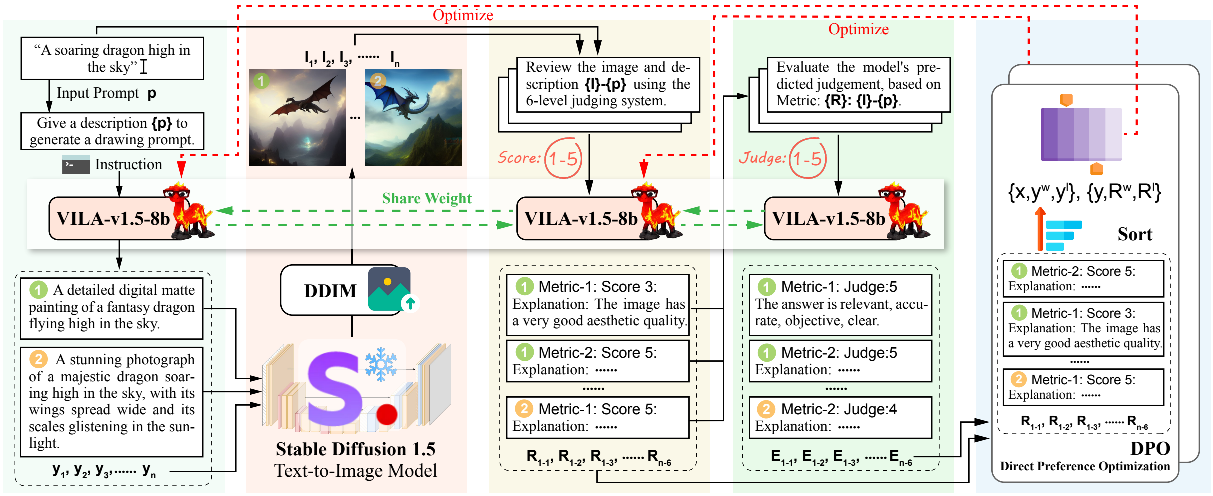
$$p_\theta(\mathbf{r}|(\mathbf{x}, \mathbf{y})) = \prod_{k=1}^{l} (r_k | (\mathbf{x}, \mathbf{y}), \mathbf{r}_{<k}) \quad (4)$$

where $\mathbf{r}_{<1}$ is usually null or a start token and $\mathbf{r}_{<k} = [r_1, r_2, .., r_{l-1}]$, $k \in \{2, ..., l\}$.

2.3 Related Work

Prompt Engineering. Hao et al. (2024) propose a prompt adaption framework for prompt engineering. To implement reinforcement learning fine-tuning, a reward function for image aesthetics and alignment is defined. Bestprompt (Pavlichenko and Ustalov, 2023) is proposed to detect keywords by genetic algorithm and then form prompts to obtain the better aesthetics of images. Beautiful-Prompt (Cao et al., 2023) first trains two reward models: Aesthetics and PickScore with a large dataset and then optimizes the language model using PPO. NeuroPrompt (Rosenman et al., 2023) utilizes constrained text decoding with a pre-trained language model to produce prompts. (Datta et al., 2023) is proposed as a Prompt Expansion framework to improve the diversity in text-to-image generation. Inspired by visual language modeling, recently, more and more methods (Li et al., 2022; Liu et al., 2024; Chen et al., 2024a; Yang et al., 2023; Yu et al., 2023) attempt to use a prior knowledge of LVLM to analyze images. The vision-language model is a large multimodal model, which bridges the gap between language and images. CLIP (Radford et al., 2021) is trained with large-scale paired text and images using contrastive learning. In this way, the pre-trained LVLMs capture rich vision-language correspondence knowledge. ALIGN (Jia et al., 2021) scales up the training process, using the larger images-text pairs but noisy data. Recently, with the great success of Large Language Models (LLMs), some work has been devoted to enabling LLMs to use image inputs. OpenFlamingo (Awadalla et al., 2023) and LLaMA-Adapter (Zhang et al., 2023) construct multimodal models based on the best LLM LlaMA (Touvron et al., 2023). To further improve the model's instruction-following abilities, LlaVA (Liu et al., 2024) employs visual instruction tuning that yields promising results. ViLA (Lin et al., 2024) achieves multi-image reasoning through a better training strategy.

Text-to-Image generation. Text-to-image models usually refer to the generative model which synthesizes an image from a given text. Earlier work, GAN (Reed et al., 2016; Tao et al., 2022) and VAE (Ramesh et al., 2021; Ding et al., 2021) have been extensively studied in this field. Recently, the diffusion-based models (Rombach et al., 2022; Gu et al., 2022) further improve the quality of


Optimize
"A soaring dragon high in the sky"
Input Prompt p
Give a description {p} to generate a drawing prompt.
Instruction
VILA-v1.5-8b
1 A detailed digital matte painting of a fantasy dragon flying high in the sky.
2 A stunning photograph of a majestic dragon soaring high in the sky, with its wings spread wide and its scales glistening in the sunlight.
$y_1, y_2, y_3, \ldots\ldots y_n$
$I_1, I_2, I_3, \ldots\ldots I_n$
Share Weight
DDIM
Stable Diffusion 1.5
Text-to-Image Model
Review the image and description {I}-{p} using the 6-level judging system.
Score: 1-5
VILA-v1.5-8b
1 Metric-1: Score 3: Explanation: The image has a very good aesthetic quality.
1 Metric-2: Score 5: Explanation: ......
2 Metric-1: Score 5: Explanation: ......
$R_{1\text{-}1}, R_{1\text{-}2}, R_{1\text{-}3}, \ldots\ldots R_{n\text{-}6}$
Optimize
Evaluate the model's predicted judgement, based on Metric: {R}: {I}-{p}.
Judge: 1-5
VILA-v1.5-8b
1 Metric-1: Judge:5 The answer is relevant, accurate, objective, clear.
1 Metric-2: Judge:5 Explanation: ......
2 Metric-2: Judge:4 Explanation: ......
$E_{1\text{-}1}, E_{1\text{-}2}, E_{1\text{-}3}, \ldots\ldots E_{n\text{-}6}$
$\{x,y^w,y^l\}, \{y,R^w,R^l\}$
Sort
1 Metric-2: Score 5: Explanation: ......
1 Metric-1: Score 3: Explanation: The image has a very good aesthetic quality.
2 Metric-1: Score 5: Explanation: ......
$R_{1\text{-}1}, R_{1\text{-}2}, R_{1\text{-}3}, \ldots\ldots R_{n\text{-}6}$
DPO
Direct Preference Optimization

(a) Prompt Rewriter (b) Diffusion Model (c) Image Evaluator (d) Response Judge (e) Optimization

Figure 3: **The overall framework in our prompt optimizing framework.** It involves five steps, arranged from left to right, **(a) Prompt Rewriter** sample multiple candidates **y**, **(b) Diffusion Model** generates images from the candidates, **(c) Image Evaluator** act as image evaluate models, to generate image evaluate responses **R**, **(d) Response Judge** act as judge models to judge the response from evaluator and get response evaluation **E**, and **(e) Optimization** with the response from evaluator and judge, and then update the LVLM.

text-to-image generation. Given the generated output is no longer a serious concern, exploring how to optimize the prompt to maximize the potential of generative models becomes the primary object of our research. Several works (Hao et al., 2024; Cao et al., 2023) have been put into unlocking the maximum potential of text-to-image models by optimizing prompts using large language models. Besides, some reinforcement learning-based methods have gained consistent improvements in changing the output preferences of large language models. Reinforcement Learning from Human Feedback (RLHF) (Yuan et al., 2024a) yields promising results in various NLP tasks (Christiano et al., 2017; Ibarz et al., 2018; Stiennon et al., 2020; Jaques et al., 2019), at the cost of collecting large amounts of human feedback. Proximal Policy Optimization (PPO) (Schulman et al., 2017) is optimized according to the reward signal rather than human feedback and remains stable by approximating the old and new strategies. With these challenges in mind, Direct Preference Optimization (DPO) (Rafailov et al., 2024) avoids the drawback of large amounts of human-annotated data to train reward models by using its language model as a reward model. The different strategies to provide rewards are shown in Figure 2.

3 Methodology

In this section, we first describe the pipeline of dataset construction (Section 3.1), and how to initialize LVLM with both prompt optimization and preference assessing capabilities (Section 3.2) and then detail self-rewarding training algorithms (Section 3.3). Next, we introduce the use of AI feedback for RL in self-rewarding training (Section 3.4) and how to transform LVLM into a reward model by LVLM-as-a-judge (Section 3.5).

3.1 Dataset Construction Pipeline

Prompt rewrite data. To equip the LVLM with initial rewriting capabilities, rewrite prompts data is first given to train in the supervised fine-tuning (SFT) manner. This type of data is presented in pairs, i.e., {raw prompt, rewritten prompt}, to guide the LVLM to rewrite prompt. Prompt rewrite data contains 104,487 prompt pairs in (Cao et al., 2023).

Evaluation data. The LVLM specializes in image understanding and analyzing (Zhu et al., 2024; Li et al., 2022). However, image aesthetics is still a subjective understanding and not easy for models. Although unnecessary, LVLM appears more difficult to score images and construct preference pairs without fine-tuning on evaluation data (Yuan et al., 2024b). Therefore, evaluation data is constructed to offer the model some examples to understand the aesthetics. The prompt data in the previous paragraph is used to generate images by diffusion model (Rombach et al., 2022). These images are then scored using the prompt template in the LVLM-as-a-judge way (Zheng et al., 2024; Chen et al., 2024a). We use a subset of the aesthetics training set and the Pickapic-v1 training set. The paired images and evaluation prompt are both fed into VILA-40B as well as the facts to obtain the response as truth response for training. We add a truth cue to the evaluation prompt, if the VILA-40B's scoring of the two images is consistent with the facts, the evaluation response is retained.

Judgment data. To improve the LVLM's evaluation capabilities for responses (Zheng et al., 2024; Zhu et al., 2023), we also create a dataset for the model to understand which responses to aesthetics or alignment are more reasonable. Specifically, we construct data that assign higher scores to reasonable responses and lower scores to irrelevant responses, which is some kind of confidence. To improve the evaluation ability of LVLM, we create a new dataset to evaluate the model's responses to aesthetics and PickScore of the images. Similarly, we feed the standard (i.e., VILA-40B's response) and irrelevant answers into VILA-40B, and if the model has a higher score in the standard answer than in the irrelevant answer, the answer remains in the dataset. A common observation is that most images generated from rewritten prompts are typically assigned scores of 3 or 4. Therefore, we also include some raw prompts to maintain a more uniform score distribution.

3.2 Initialization LVLM Capability of Rewriting and Assessing Preference

After we obtain the initial dataset, supervised fine-tuning (SFT) is employed to initiate the model with the ability of image evaluation and response judgment. Therefore, we consider the input $\mathbf{x}$ for a specific task to be derived from the distribution $q(\cdot)$. Meanwhile, the probability distribution in the SFT training data can be represented as $p_{data}(\cdot|\mathbf{x})$. The object function of SFT can be represented as:

$$\mathcal{L}_{SFT}(\theta) = -\mathbb{E}_{\mathbf{x}\sim q(\cdot),\mathbf{y}\sim p_{data}(\cdot|\mathbf{x})}[\log p_\theta(\mathbf{y}|\mathbf{x})] \quad (5)$$

Since the SFT training data has high-quality labeled responses $\mathbf{y}$, the main purpose of this object function is to approximate the model's predictive distribution $p_\theta(\mathbf{y}|\mathbf{x})$ to the target $p_{data}(\mathbf{y}|\mathbf{x})$.

3.3 Training with Self-Rewarding

The main objective of the self-rewarding model is to complete iterative training with limited amounts of human-labeled data on the pre-trained model. During an iteration, the model generates a solution and a judgment on this solution. Unlike other methods that use a fixed reward model (Ouyang et al., 2022), the self-rewarding model uses the model of the current iteration to generate the rewards. Therefore, the model can both improve its generative capabilities when acting as a generative model and get its boost when acting as a rewarding model, since the corresponding responses are generated by the same mechanism (Yuan et al., 2024b).

The whole training process is concluded as follows: the model first starts with pre-trained LVLMs (denoted as LVLM_{init}), training with prompt rewrite data and evaluation data in SFT manner, resulting in LVLM_{SFT}. To further enhance the model's performance, we sample multiple candidates from the raw prompt and construct preference pairs based on the LVLM scores, and then the model is trained in the DPO manner. This training process is iterated M times, yielding models denoted as $\{\text{LVLM}_{\text{DPO}_1}, \text{LVLM}_{\text{DPO}_2}, \cdots, \text{LVLM}_{\text{DPO}_M}\}$. The bottleneck of prompt engineering can be summarized into two aspects: 1) it is difficult to align with human preference and thus a large number of human-labeled data is required; 2) the whole training process is too complex, typically requiring a larger trained reward model. Therefore, we propose a self-rewarding prompt rewriting model to obtain better performance through AI self-feedback and an evolving training process. The overall framework is shown in Fig. 3. First, the model acts as a solver sampling multiple candidate answers from raw prompts. These candidates are then fed into a fixed text-to-image model to generate the corresponding images. Again, the previous solver is used as the reward model to generate responses, for scoring these images generated from candidates. The details of scoring are discussed below. Finally, preference pairs are constructed for model DPO training.

3.3.1 Training on Rewriting

We consider an LVLM to be parameterized by θ and denoted by p_θ. The model takes a prompt as sequence $\mathbf{x} = [x_1, x_2, ..., x_n]$ to generate the corresponding response $\mathbf{y} = [y_1, y_2, ..., y_m]$. The response $\mathbf{y}$ is thus regarded as a sample from the conditional probability distribution $p_\theta(\cdot|\mathbf{x})$. Specifically, x_i and y_j represent the tokens from the same predetermined vocabulary within the sequences $\mathbf{x}$ and $\mathbf{y}$, respectively. To generate the y_j for a given position, the auto-regressive model p_θ exploits the previously generated tokens to generate subsequent tokens up to the maximum length or the end token. Therefore, the conditional probability distribution

$p_\theta(\mathbf{y}|\mathbf{x})$ can be expressed as:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{j=1}^{m} (y_j|\mathbf{x}, \mathbf{y}_{<j}) \tag{6}$$

where $\mathbf{y}_{<1}$ is usually null or a start token and $\mathbf{y}_{<j} = [y_1, y_2, .., y_{j-1}], j \in \{2, ..., m\}$.

3.3.2 Training on Preference Assessing

Beyond the model training on prompt rewriting, we also perform preference assessing training on the same model to improve its assessing ability. Similar to Sec. 3.3.1, the model generates the judgment on the response **R** for a response **y** on prompt and the generated image. The conditional probability distribution can be defined as:

$$p_\theta(\mathbf{R}|\mathbf{y}) = \prod_{k=1}^{n} (R_j|\mathbf{y}, \mathbf{R}_{<k}) \tag{7}$$

where $\mathbf{y}_{<1}$ is usually null or a start token and $\mathbf{R}_{<k} = [R_1, R_2, .., R_{k-1}], j \in \{2, ..., n\}$.

Therefore, RL fine-tuning in self-rewarding can be optimized with the loss function, i.e.,

$$\begin{aligned} \mathcal{L}_{RL}(\theta_{t+1}) = {} & \mathbb{E}_{\mathbf{x}\sim q(\cdot), \mathbf{y}\sim p_{\theta_t}, \mathbf{r}\sim p_{\theta_t}(\cdot|(\mathbf{x},\mathbf{y}))}[r(\mathbf{x}, \mathbf{y})] \\ & - \lambda \mathbb{E}_{\mathbf{x}\sim q(\cdot)} \mathrm{KL}(p_\theta(\cdot|\mathbf{x}) || p_{ref}(\cdot|\mathbf{x})) \\ & + \mathbb{E}_{\mathbf{y}\sim q(\cdot), \mathbf{R}\sim p_{\theta_t}, \mathbf{E}\sim p_{\theta_t}(\cdot|(\mathbf{y},\mathbf{R}))}[E(\mathbf{y}, \mathbf{R})] \\ & - \lambda \mathbb{E}_{\mathbf{y}\sim q(\cdot)} \mathrm{KL}(p_\theta(\cdot|\mathbf{y}) || p_{ref} \end{aligned} \tag{8}$$

where θ_t denotes the t-th parameters of the model, the $r(\mathbf{x}, \mathbf{y})$ and the $E(\mathbf{y}, \mathbf{R})$ denote the reward function of the prompt x and the response y, respectively. Since the self-reward model is an iterative model, in which the iterative process results in a series of models with different weights of the same structure. For better understanding, the parameter of the result model by Equation (5) is denoted as θ_0, while the parameters of the result models optimized through Equation (8) are denoted as $\{\theta_1, \theta_2, \cdots\}$. As the model's parameters continue to be optimized for the ability to follow instructions, so does the reward ability, thereby facilitating self-improvement.

3.4 RL from AI Feedback

AI Feedback. One of the main investigations of our work is how to use prior knowledge of large models to guide text-to-image models. Unlike time-consuming and labour-intensive human feedback training, AI feedback (Lee et al., 2023) training can provide reward signals for a given task through its own knowledge. This approach usually requires two models, one acting as a solver of the downstream task, and one acting as a judge of the solver. It is feasible to apply an external language model as a judge (or reward model) or to use only itself (i.e. a model that acts as both a solver and a judge) to achieve self-improvement in a specific task.

Therefore, preference pairs ⟨ raw prompt x, winner prompt y^w, loser prompt y^l ⟩ need to be constructed to train the model. The model first generates different N candidates from randomly selected raw prompts (in the previous data). After generating images using a fixed text-to-image model, the images and evaluation prompts are input into the model for scoring. Next, the prompts corresponding to the highest and lowest-scored images are treated as winners and losers, respectively. In addition, the image evaluation responses are judged by LVLM and the pairwise data for DPO are also constructed. The model is then tuned with DPO (Rafailov et al., 2024).

DPO. Assuming access to a static dataset of comparisons $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i^w, \mathbf{y}_i^l\}$, which is sampled from p^*. The optimal RLHF policy π^* under the Bradley-Terry model satisfies the preference model:

$$\begin{aligned} p^*(\mathbf{y}^1 \succ \mathbf{y}^2|\mathbf{x}) = \Big[1 + \exp\left(\lambda \log \frac{\pi^*(\mathbf{y}^2|\mathbf{x})}{\pi_{ref}(\mathbf{y}^2|\mathbf{x})}\right) \\ - \lambda \log \frac{\pi^*(\mathbf{y}^1|\mathbf{x})}{\pi_{ref}(\mathbf{y}^1|\mathbf{x})}) \Big]^{-1} \end{aligned} \tag{9}$$

Given a preference pair $\langle \mathbf{x}, \mathbf{y}^w, \mathbf{y}^l \rangle$, the object function of DPO is to seek a maximum likelihood of the parameterized policy π_θ by reference model π_{ref}.

$$\begin{aligned} & \mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) \\ & = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}^w, \mathbf{y}^l)\sim\mathcal{D}} \Big[\log \sigma \big(\Delta_\lambda(\mathbf{y}^w|\mathbf{x}) - \Delta_\lambda(\mathbf{y}^l|\mathbf{x}) \big) \Big], \\ & \text{where } \Delta_\lambda(y|x) = \lambda \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}. \end{aligned} \tag{10}$$

The λ is a parameter controlling the deviation from the base reference policy π_{ref}.

3.5 LVLM-as-a-judge for Aesthetics and Rewarding

To improve the text-to-image models, we consider this from two perspectives: **the aesthetics of the generated image** and **the capability to follow instructions**, respectively. Hence, to empower the model in generating appropriate reward signals for the images, we establish a judging template comprising aesthetic score, pick score and alignment score. This enables the LVLM to evaluate the generated images effectively. Details of prompts for LVLM-as-a-judge can be found in Appendix.

Aesthetic Score. Aesthetics are assessed by prompting the model to consider aspects such as composition, color, lighting, and visual appeal. To facilitate judgment-making, we design a grading system that allows the model to attribute a certain grade to the generated images, thus obtaining the corresponding score. For each evaluation, the model is asked to provide both the score and a brief explanation, employing a chain-of-thought reasoning approach.

PickScore (Kirstain et al., 2024). PickScore is an important metric to measure human preferences. We consider whether the image represents the given text in a way that is favored by humans. Nonetheless, it is difficult for the model to understand human preferences directly, so we included evaluation data (details in Section 3.2) for model initialization.

Relevance Score. The judgment template for the relevance score uses the same hierarchical form as the aesthetic score. The key difference is that it evaluates the model's attention to the user instructions for text-to-image generation in three areas: presence of the object, accurate count of the object, and correct relationships between objects.

4 Experiments

4.1 Settings

Experiments are conducted on the public text-to-image model Stable Diffusion v1.5, and the denoising steps are set to 20 to accelerate the image sampling. The base model we employ in this paper is the smaller model VILA-v1.5-8b (Lin et al., 2024). For SFT, we use AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.95$), with a batch size of 16 and a weight decay of 0.1. We use an initial learning rate of 2e-5, with a linear warm-up and cosine decay schedule. To improve the diversity of the candidates, we sample $N = 8$ candidates with temperature $T = 0.9, p = 0.9$ from one raw prompt. When validating these candidates with the reward model, we utilize a temperature $T = 0.9, p = 0.9$, sampling three times and averaging the scores to determine the final score for further DPO training. The overall score is calculated by summing up the aesthetics score, pick Score and the relevance score. The highest one and the lowest one are kept as winners and losers but discarded if they have the same score. We perform two DPO iterations and the size of each prompt rewrite training data is 10k and 20k, respectively. Besides, 10k judgment data is employed to improve image evaluation. In the stage of DPO, we employ AdamW with an initial learning rate of 1e-5, 5e-6 without weight decay. The batch size is set to 32 and β in DPO is set to 0.1. The model is trained for 4 epochs at each iteration. All experiments are conducted on $4\times$ NVIDIA A800 80G GPUs. For evaluation, we adopt beam search (Vijayakumar et al., 2016) with a beam size of 4 and a length penalty of 1.0.

Method	$\mathbb{D}_{rl}$	**PickScore**	**Aes.**	**CLIP**
Original	-	20.74	5.50	**0.27**
MagicPrompt (Santana, 2022)	-	20.11	5.79	0.22
ChatGPT (OpenAI, 2023)	-	20.73	5.92	0.25
Beautiful-Prompt (Cao et al., 2023)	40k	20.84	6.52	0.24
Our Method				
LVLM_{SFT}	-	20.79	5.95	0.25
$\text{LVLM}_{\text{DPO}_1}$	10k	20.81	6.31	0.24
$\text{LVLM}_{\text{DPO}_2}$	20k	**20.86**	**6.59**	0.24

Table 1: Evaluation of the aesthetic score and CLIP score on Beautiful-Prompt test set. "$\mathbb{D}_{rl}$" means the size of the train set we used in reinforcement learning. "Aes." and "CLIP" mean the aesthetic and CLIP scores, respectively.

Method	**Type**	**Aes.**	**CLIP**
Original†	Human	5.47	**0.28**
Human Engineered Prompt†	Human	5.87	0.26
NeuroPrompts(Rosenman et al., 2023)	AI	6.27	-
Promptist(Hao et al., 2024)	AI	6.26	0.26
$\text{LVLM}_{\text{DPO}_2}$ (Ours)	AI	**6.57**	0.26

Table 2: Evaluation of the aesthetic score and CLIP score on DiffusionDB. Values marked with † from (Hao et al., 2024).

4.2 Comparative Methods

We compare our method with the following approaches: MagicPrompt (Santana, 2022), ChatGPT (OpenAI, 2023), Beautiful-Prompt (Cao et al., 2023), NeuroPrompts (Rosenman et al., 2023) and Promptist (Hao et al., 2024). It is worth noting that, we prompt ChatGPT (OpenAI, 2023) to generate an expansion of the user-provided prompt for generative models. Other models generate results using their open-source weights. The "Human Engineered Prompt" refers to the prompts written by humans, while its simplified version of these prompts is used as the original prompt. In real-world usage, users only need to provide simple prompt to obtain high-quality results.

4.3 Results

The model is validated on two datasets, Beautiful-Prompt (Cao et al., 2023) and DiffusionDB (Wang et al., 2023) (100k), respectively. As shown in Table 1, our method outperforms other methods. This

Pick	Aes.	Align	Reward	PickScore	Aes.	CLIP
✓	✓	✓	Self	**20.86**	6.59	**0.24**
	✓	✓	Self	20.73	6.59	**0.24**
✓	✓		Self	20.79	**6.61**	0.22
✓		✓	Self	20.76	6.49	**0.24**
✓	✓	✓	Fixed	20.73	6.26	**0.24**

Table 3: Ablation study on different prompts and reward mode. Checkmark is enabled prompts in image evaluation.

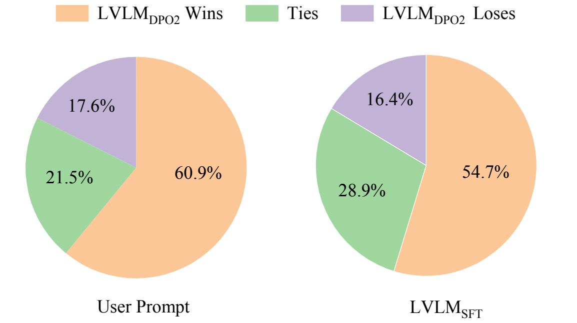


Figure 4: Human evaluation results. The result of LVLM$_{DPO_2}$ are more preferred by human compared with the result of User Prompt and LVLM$_{SFT}$.

performance gap is more evident on larger datasets DiffusionDB (100k). As can be seen in Table 2, our method achieves a performance gain compared to other methods. Compared to the most competitive method Beautiful-Prompt, our method achieves a 0.1 improvement in the Aesthetic Score, while maintaining no degradation in the PickScore and CLIP score. However, on the larger DiffusionDB test set, our method achieves a 0.3 improvement compared to the Neuroprompt (Rosenman et al., 2023) and Promptist (Hao et al., 2024).

4.4 Ablation Study

We perform ablation experiments from two perspectives: how the prompt affects the LVLM as a reward model and how self-rewarding training improves the model. First, as can be seen in Table 3, the model without Pick Prompt for rewarding received a high aesthetic score but leads a significant drop in PickScore. It also suggests that some shifts exist between human preferences and image aesthetics. In addition, it can be observed a reduction of the aesthetic score and CLIP score when missing the aesthetic prompt or relevance prompt. Second, we compare the performance between the model using self-rewarding and fixed-rewarding. Self-rewarding outperforms the fixed-rewarding methods. More discussion about the self-rewarding method can be found in the Appendix D.

4.5 Human Evaluation

To better compare the existing methods with our model from the human perspective, a human study is conducted. We randomly select prompts and obtain the optimized prompt through different models, and then the images are generated. Then, 20 volunteers are asked to rank the different images. Each volunteer is presented with two images at a time and asked to select the one they find more appealing. The volunteers in this study are randomly selected from a pool of individuals with diverse educational backgrounds. The result is shown in Figure 4. It can be observed that the images obtained through the optimized prompt generated by our model are selected most frequently by the volunteers. And the DPO model is better than the SFT baseline, which achieves about 55% win rate.

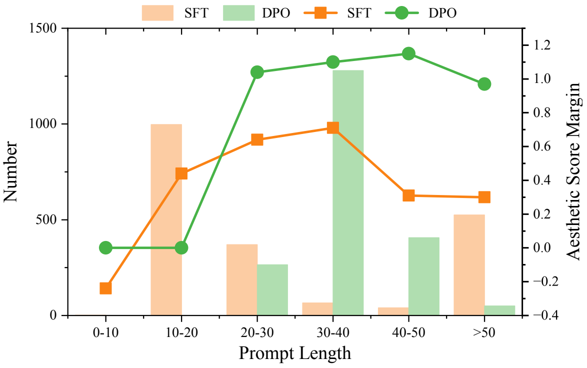


Figure 5: Generated prompt length (bar graph) and aesthetic score improvement (line graph) compared with raw prompt.

# Training Data			
Stage	Beautiful-Prompt	Promptist	Ours
SFT	143k	360k	156k
Rewrite	143k	360k	104k
Evaluation	-	-	32k
Judgment	-	-	20k
RM	7M	-	-
RL	40k	90k	40k
Rewrite	40k	90k	{10k, 20k}
Evaluation	-	-	10k

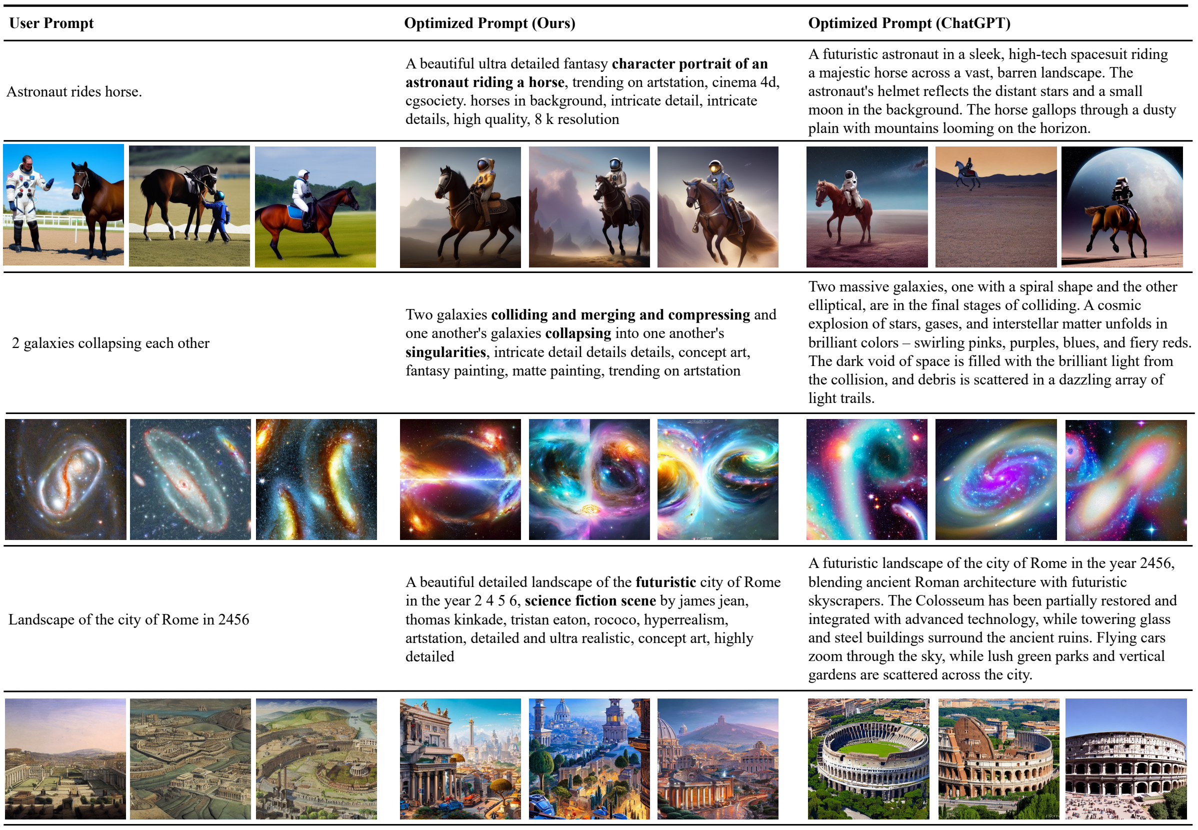
Table 4: Training set number on different stages of Beautiful-Prompt(Cao et al., 2023), Promptist(Hao et al., 2024), and ours. **RM** means the training of the reward model. Our method requires addition data on image evaluation and response judgment for rewarding.

Method	LVLM$_{init}$	LVLM$_{\mathrm{SFT}}$	LVLM$_{\mathrm{DPO}_1}$
Test-Set Accuracy	48.0%	66.1%	66.4%

Table 5: Evaluation on Pickapic-v1 test set. Our self-rewarding model realizes the ability to self-improvement

4.6 Further Discussion

To further understand how the prompt optimization affects the original prompt, we show some analysis in Figure 5. Considering the length of prompts has a significant relationship with the quality of generation, we count the length of optimized prompts



User Prompt	Optimized Prompt (Ours)	Optimized Prompt (ChatGPT)
Astronaut rides horse.	A beautiful ultra detailed fantasy **character portrait of an astronaut riding a horse**, trending on artstation, cinema 4d, cgsociety. horses in background, intricate detail, intricate details, high quality, 8 k resolution	A futuristic astronaut in a sleek, high-tech spacesuit riding a majestic horse across a vast, barren landscape. The astronaut's helmet reflects the distant stars and a small moon in the background. The horse gallops through a dusty plain with mountains looming on the horizon.
2 galaxies collapsing each other	Two galaxies **colliding and merging and compressing** and one another's galaxies **collapsing** into one another's **singularities**, intricate detail details details, concept art, fantasy painting, matte painting, trending on artstation	Two massive galaxies, one with a spiral shape and the other elliptical, are in the final stages of colliding. A cosmic explosion of stars, gases, and interstellar matter unfolds in brilliant colors – swirling pinks, purples, blues, and fiery reds. The dark void of space is filled with the brilliant light from the collision, and debris is scattered in a dazzling array of light trails.
Landscape of the city of Rome in 2456	A beautiful detailed landscape of the **futuristic** city of Rome in the year 2 4 5 6, **science fiction scene** by james jean, thomas kinkade, tristan eaton, rococo, hyperrealism, artstation, detailed and ultra realistic, concept art, highly detailed	A futuristic landscape of the city of Rome in the year 2456, blending ancient Roman architecture with futuristic skyscrapers. The Colosseum has been partially restored and integrated with advanced technology, while towering glass and steel buildings surround the ancient ruins. Flying cars zoom through the sky, while lush green parks and vertical gardens are scattered across the city.

Figure 6: The generated images with the optimized prompts using our method

of LVLM$_{SFT}$ and LVLM$_{DPO_2}$ and their aesthetic score margins compared to the original prompt. An immediate observation is that the length of most optimized prompts is between 30 and 40. Besides, aesthetic scores gradually increased with the length. Besides, we also present the size of the training set in Table 4. In SFT stage, we use prompt rewrite data (104k), evaluation data (32k) and judgment data (20k), while in RL fine-tuning stage, we use prompt data 10k, 20k and new evaluation data 10k sampled from LVLM. Note that we just expanded the dataset after each iteration, whereas 10k prompt data is the subset of 20k. This means only 20k training data are needed to train this model, while Beautiful-Prompt and Promptist employ the larger training set, i.e., 40k and 90k, respectively. To better exemplify the ability of our model to evaluate by self-rewarding, we show the accuracy of the test set in Pickapic-v1 after SFT and DPO training. As shown in Table 5, our model realizes an increase in evaluation capacity.

4.7 Quantitative Results

We present more visual results in Figure 6, it can be observed that our method rewrites the main content appropriately, and describes the detail, e.g., art style, light, etc. The images generated by these optimized prompts have better aesthetics than the images generated from the original prompts, which are bland visually. In addition, some incorrect prompts can also be corrected. For example, the prompt *“Astronaut rides horse"* mislead the text-to-image models to generate an astronaut standing beside a horse. However, our approach rewrites it as *“...an astronaut riding a horse..."* to make the image more in line with the user’s intention. In addition, the model expands some descriptions on some objects, like *“futuristic""* relative to the *“in 2456”*. More results are in the Appendix C.

5 Conclusion

In this work, we propose a novel method to optimize prompts for text-to-image models, which can be used to fill the gap between laymen and experts when using a generative model. We first transform the LVLM into a reward model so that it can judge the aesthetics and alignment of the images. In so doing, we can perform AI feedback rather than human feedback. Then, to gain the performance boost based on limited data, we employ self-rewarding training for LVLM. Our model achieves self-improvement through an iterative training approach. Experiments on two datasets show that our method outperforms the other methods.

Limitations

The primary limitation of our work is the inability to utilize larger parameter models due to the significant computational resources required. Training and fine-tuning models with billions of parameters, such as VILA-40B, demand substantial GPU memory and processing power, which are often constrained by available hardware. Deploying large-parameter models in real-world applications presents considerable challenges.

References

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–11.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. 2023. Prompt expansion for adaptive text-to-image generation. *arXiv preprint arXiv:2312.16720*.

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

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

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2024. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in Neural Information Processing Systems*, 31.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2024. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M Ni, PengChuan Zhang, and Lei Zhang. 2022. Vision-language intelligence: Tasks, representation learning, and large models. *arXiv preprint arXiv:2203.01922*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

OpenAI. 2023. Chatgpt (mar 14 version) [large language model]. *https://chat.openai.com/chat*.

Jonas Oppenlaender. 2023. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, pages 1–14.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Nikita Pavlichenko and Dmitry Ustalov. 2023. Best prompts for text-to-image models and how to find them. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2067–2071.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Shachar Rosenman, Vasudev Lal, and Phillip Howard. 2023. Neuroprompts: An adaptive framework to optimize prompts for text-to-image generation. *arXiv preprint arXiv:2311.12229*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Gustavo Santana. 2022. Magicprompt - stable diffusion.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525.

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

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2023. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2024a. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024b. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Comparisons with the common tags

To figure out what word improves the prompt, we count the most frequent words. Evidently, *"fantasy"* and *"intricate"* are popular in the model, which even appears more than once in the same sentence.

To demonstrate that the performance improvement of the prompt rewriting method is not the result of adding some fixed words, we select the most frequent words and randomly combine them into six tags. Then, we combine it with the original prompt and calculated the scores. The tags are shown in Tab. 7 and the results are present in Tab. 8.

Content	Frequency
fantasy	2,291
intricate	2,146
portrait	1,604
beautiful	1,211
highly detailed	701
realistic	596
high quality	389
elegant	381
illustration	121

Table 6: Analysis of the most frequent words in the optimized prompts in the beautiful-prompt test set.

Tag	Content
1	artstation, highly detailed, elegant
2	8k, trending on artstation, concept art
3	digital painting, intricate, fantasy
4	illustration, smooth, fantasy
5	portrait, beautiful, illustration
6	realistic, dramatic, high quality

Table 7: Combinations of common tags

Metric	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6	Ours
PickScore	20.81	20.73	20.68	20.59	20.69	20.82	20.86
Aes. Score	5.80	5.75	5.92	5.67	5.75	5.59	6.59

Table 8: The result using different groups of common tags.

B Further Discussion

An essential mechanism in self-rewarding model is sampling multiple outputs and then judging them to construct preference pairs for RL fine-tuning. Thus, the diversity of candidates greatly affects the performance of self-rewarding, which can be demonstrated from Tab. 9. It is essential to enlarge the margin between winners and losers with more candidates, which facilitates RL fine-tuning.

Method	Reward	#Candidates	PickScore	Aesthetic	CLIP
LVLM_{SFT}	-	-	20.79	5.95	**0.25**
$\text{LVLM}_{\text{DPO}_1}$	LVLM_{SFT}	2	20.65	6.03	0.23
$\text{LVLM}_{\text{DPO}_1}$	LVLM_{SFT}	4	20.78	6.16	0.23
$\text{LVLM}_{\text{DPO}_1}$	LVLM_{SFT}	8	20.81	6.31	0.24

Table 9: Number of candidates

Method	PickScore	Aes.	CLIP
LVLM_{SFT}	20.79	5.95	**0.25**
$\text{LVLM}_{\text{DPO}_1}$	20.81	6.31	0.24
$\text{LVLM}_{\text{DPO}_2}$	**20.86**	6.59	0.24
$\text{LVLM}_{\text{DPO}_3}$	**20.86**	**6.60**	0.24

Table 10: More iteration training.

In addition, the self-reward model is at risk of bias or overfitting. There are many factors that may contribute to model collapse, including data quality and scale, the inherent capabilities of the language model, biases in the generative model, and the number of candidate samples. To demonstrate the effectiveness of our iterative training, we show the results of one more round of training. We used an additional 5,000 data for one more training. As shown in Tab. 10, the third iteration does not cause a significant performance drop.

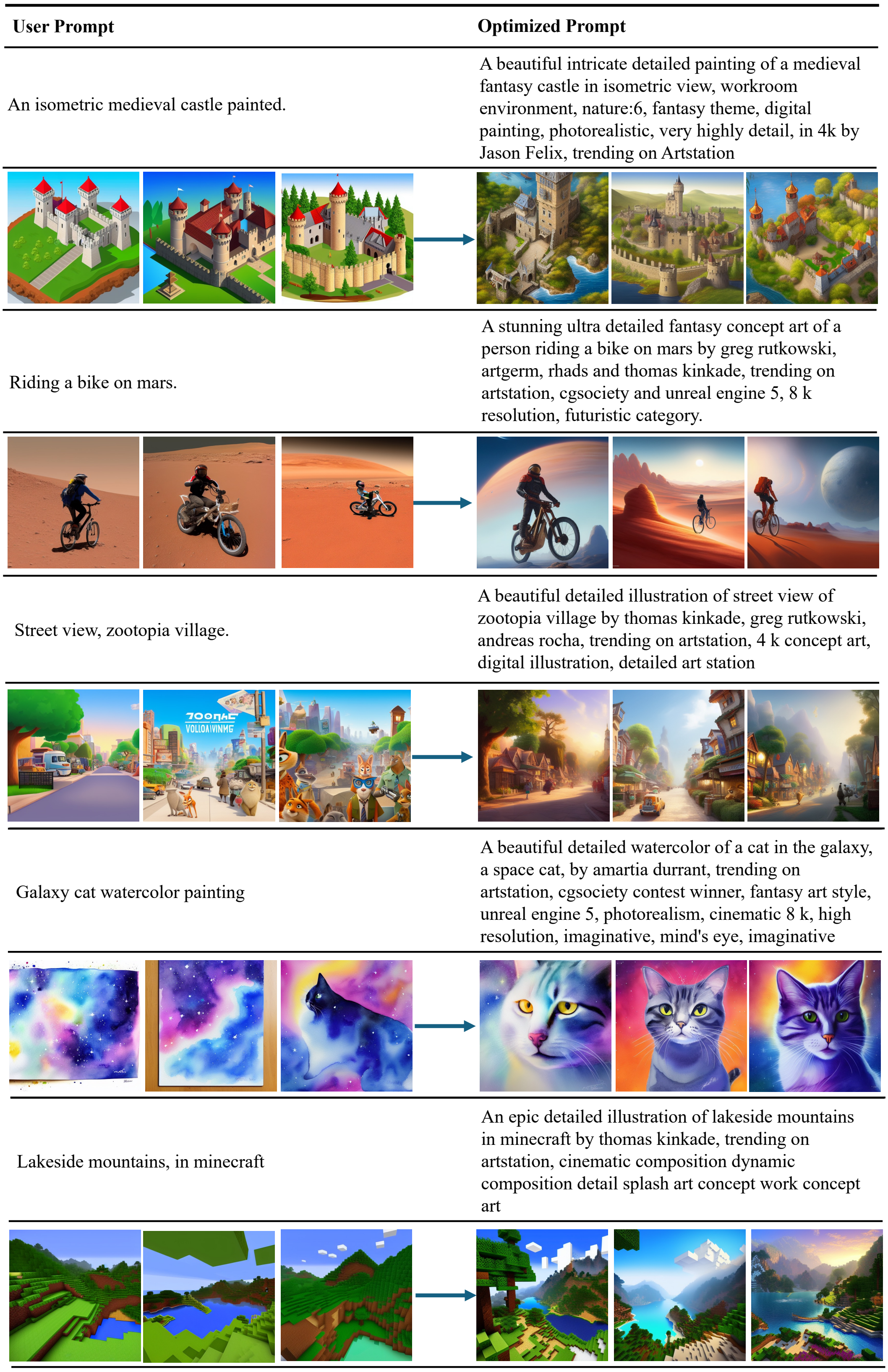
C More Quantitative Results

We present more visual results between the images generated with different prompts. As shown in Fig. 7, the optimized prompts result in more pleasing images. A more intuitive observation is that the flat, uninteresting view in Minecraft and the more aesthetically pleasing, more detailed view are represented by the optimized prompt before and after optimization, respectively. In addition, the modified prompt has stronger alignment capabilities. For example, the prompt "*Riding a bike on mars*" is amended to "*... a person riding a bike on mars ...*", and the prompt "*Galaxy cat*" is added with the description "*... a space cat ...*".

D Prompts for LVLM-as-a-judge

We present the prompt to transform the LVLM into the reward model. As mentioned in Sec. 3.5, we employ the rating system for image aesthetics, human preferences, and alignment. Note that we input the generated image by original prompt and the image by candidate prompt here to facilitate scoring of the model on the same benchmarks.

For responses judgment, we input two responses to the model that are about how the model evaluated the image in the previous step.



User Prompt	Optimized Prompt
An isometric medieval castle painted.	A beautiful intricate detailed painting of a medieval fantasy castle in isometric view, workroom environment, nature:6, fantasy theme, digital painting, photorealistic, very highly detail, in 4k by Jason Felix, trending on Artstation
Riding a bike on mars.	A stunning ultra detailed fantasy concept art of a person riding a bike on mars by greg rutkowski, artgerm, rhads and thomas kinkade, trending on artstation, cgsociety and unreal engine 5, 8 k resolution, futuristic category.
Street view, zootopia village.	A beautiful detailed illustration of street view of zootopia village by thomas kinkade, greg rutkowski, andreas rocha, trending on artstation, 4 k concept art, digital illustration, detailed art station
Galaxy cat watercolor painting	A beautiful detailed watercolor of a cat in the galaxy, a space cat, by amartia durrant, trending on artstation, cgsociety contest winner, fantasy art style, unreal engine 5, photorealism, cinematic 8 k, high resolution, imaginative, mind's eye, imaginative
Lakeside mountains, in minecraft	An epic detailed illustration of lakeside mountains in minecraft by thomas kinkade, trending on artstation, cinematic composition dynamic composition detail splash art concept work concept art

Figure 7: The images generated using our optimized prompt.

Aesthetic Score Prompt

Please evaluate the aesthetics of two images (“Image 1” and “Image 2”) using the 6-level judging system described below.
The two images given are independent, and should be evaluated separately and step by step, ensuring that the order in which the images were presented does not affect your judgment.

- Poor (Score: 0): The image lacks balance, composition, and visual appeal. Colors may be overly saturated or dull, causing discomfort to the viewer. Composition is chaotic, distracting, or poorly framed.

- Below Average (Score: 1): The image present minimal aesthetic appeal, even if there are inconsistencies or major flaws in composition, color, lighting, or other aesthetic elements, or make people feel disjointed or unbalanced, lacking a cohesive visual narrative.

- Average (Score: 2): The image exhibits adequate aesthetic quality contributes to the image’s visual appeal to some extent but there is room for improvement in terms of creativity or originality or some aspects of the image may feel generic or uninspired.

- Above Average (Score: 3): The image has strong aesthetic quality regardless of whether there are minor imperfections in composition, color, lighting, or other aesthetic elements may still be present but do not significantly detract from the overall aesthetic, or aesthetic choices may be subjective, with some viewers preferring different styles or approaches.

- Very Good (Score: 4): The image is of exceptional aesthetic quality and demonstrates creativity, skill, and mastery of visual elements even if there is slight room for improvement in composition, color, lighting, or other aesthetic elements.

- Excellent (Score: 5): The image is of perfect balance, harmony, and creativity in aesthetics, creating a visually compelling and impactful image.

Please provide a comprehensive explanation of your score.
Note that the score has nothing to do with image input order.

Output format:

Output for Image 1:
Score: <Your Score for Image 1>
Explanation: <detailed judgment of Score for Image 1>

Output for Image 2:
Score: <Your Score for Image 2>
Explanation: <detailed judgment of Score for Image 2>

Alignment Score Prompt

Please evaluate the alignment of two pictures (“Image 1” and “Image 2”) to the text (“Text”) using the 6- level judging system described below.
The two images given are independent, and should be evaluated separately and step by step, ensuring that the order in which the images were presented does not affect your judgment.
You need to first consider what appears in the image, then whether what is described in the text appears in the image, and finally give a score based on the system.

Judging system:

- No Match (Score: 0): The image does not contain any of the objects or elements mentioned in the text. There is no recognizable connection between the text and the image.

- Poor Match (Score: 1): The image contains one or a few of the objects mentioned in the text, but these are peripheral and do not capture the primary content or relationships described. Quantitative relationships are ignored or inaccurately represented.

- Fair Match (Score: 2): Some of the primary objects mentioned in the text are present in the image, and at least one quantitative relationship or object relationship is correctly depicted. However, several key objects or relationships are missing or inaccurately represented.

- Good Match (Score: 3): The majority of the objects mentioned in the text are present, and many of the described quantitative relationships and object relationships are accurately depicted. Minor details may be missing or slightly altered.

- Excellent Match (Score: 4): Nearly all objects described in the text are accurately represented in the image, including precise quantitative relationships and interactions between objects. Only trivial discrepancies or omissions are present, which do not significantly impact the overall accuracy.

- Perfect Match (Score: 5): The image perfectly matches the text in terms of the presence of all described objects, accurate quantitative relationships, and the exact relationships between objects. Every detail mentioned in the text is present and correctly depicted in the image.

Text: <PROMPT>

Please provide a comprehensive explanation of your score. Note that the score has nothing to do with image input order.

Output format:

Output for Image 1:
Score: <Your Score for Image 1>
Explanation: <detailed judgment of Score for Image 1>

Output for Image 2:
Score: <Your Score for Image 2>
Explanation: <detailed judgment of Score for Image 2>

Pick Score Prompt

Please evaluate how well these two images (“Image 1” and “Image 2”) generated based on the text (“Text”) are preferred by humans using the 6-level judging system described below. The two images given are independent, and should be evaluated separately and step by step, ensuring that the order in which the images were presented does not affect your judgment. In this system, ’attractiveness’ refers to the visual appeal of an image to the human in terms of color, composition, lighting, style, and detail.

Judging system:

- Poor (Score: 0): The image is repulsive or offensive, lacking any attractiveness. It is completely irrelevant to the text information and presents the text in a manner that is unpleasant or unacceptable to the audience.

- Below Average (Score: 1): The image has almost no attractiveness, and the audience is indifferent to it. Its relevance to the text information is low, and the presentation style is not attractive enough, making the audience find it rather dull.

- Average (Score: 2): The image lacks attractiveness, is ordinary and lacks visual highlights, the conveyed text information is not sufficiently clear, and the presentation style is rather ordinary, lacking novelty or appeal.

- Above Average (Score: 3): The image’s attractiveness is average, without any particular outstanding features but also not mediocre, conveying the text information and presenting the text in a generally acceptable manner, albeit not particularly outstanding.

- Very Good (Score: 4): The image is highly attractive, with good visual effects, conveying the text information basically, and presenting the text in a way that is appealing to humans, allowing the audience to understand and resonate to some extent.

- Excellent (Score: 5): The image is extremely attractive, with outstanding visual effects, clearly and accurately conveying the text, and presenting the text in a way that resonates deeply with the audience and evokes strong emotional connections.

Text: <PROMPT>

Please provide a comprehensive explanation of your score.

Note that the score has nothing to do with image input order.

Output format:

Output for Image 1:
Score: <Your Score for Image 1>
Explanation: <detailed judgment of Score for Image 1>

Output for Image 2:
Score: <Your Score for Image 2>
Explanation: <detailed judgment of Score for Image 2>

Aesthetic Judgment Prompt

You are a helpful and precise assistant for checking the quality of the answers.
Given the input:

1. Image 1 and Image 2
2. Question: {{question}}
3. Answer A: {{answer_A}}
4. Answer B: {{answer_B}}

Your task is to evaluate the model's predicted answer, based on the context provided by the images and the question. There are two image scores for each answer, and you need to include an evaluation of both outputs ("Output of Image 1" and "Output of Image 2") in each answer. Please provide a comprehensive explanation of your score, noting that your explanation should be based on the facts of the images and not be vague and uninformative.

Consider the following criteria for evaluation:

- Relevance: Does each output in the predicted answer relate to the content of each image?

- Accuracy: Does the prediction in each output accurately reflect the information given in the image without introducing factual inaccuracies?

- Objectivity: For the analysis of the images, do the two predicted outputs in each answer give approximate scores, avoiding any overestimation or underestimation?

- Clarity: Assess the clarity of the predicted answer. Look for issues such as repetition, unclear descriptions, or any grammatical errors that could hinder understanding.

- Completeness: Determine if each predicted output in answer fully covers the scope of the images. Does it leave out critical information or does it include all necessary details?

Output Format:

Output for Answer A:
Score: <an integer score of quality from 1-5>
Explanation: <detailed judgment of prediction for "Output of Image 1" and "Output of Image 2">

Output for Answer B:
Score: <an integer score of quality from 1-5>
Explanation: <detailed judgment of prediction for "Output of Image 1" and "Output of Image 2">

Pick Judgment Prompt

You are a helpful and precise assistant for checking the quality of the answers.
Given the input:

1. Image 1 and Image 2
2. Prompt: {{prompt}}
3. Question: {{question}}
4. Answer A: {{answer_A}}
5. Answer B: {{answer_B}}

Your task is to evaluate the model's predicted answer, based on the context provided by the images and the question. There are two image scores for each answer, and you need to include an evaluation of both outputs ("Output of Image 1" and "Output of Image 2") in each answer. Please provide a comprehensive explanation of your score, noting that your explanation should be based on the facts of the images and not be vague and uninformative.

Consider the following criteria for evaluation:

- Relevance: Does each outputs in the predicted answer relate to the content of each image?

- Accuracy: Does the prediction in each output accurately reflect the information given in the image without introducing factual inaccuracies?

- Objectivity: For the analysis of the images, do the two predicted outputs in each answer give approximate scores, avoiding any overestimation or underestimation?

- Clarity: Assess the clarity of the predicted answer. Look for issues such as repetition, unclear descriptions, or any grammatical errors that could hinder understanding.

- Completeness: Determine if each predicted output in answer fully covers the scope of the images. Does it leave out critical information or does it include all necessary details?

Output Format:

Output for Answer A:
Score: <an integer score of quality from 1-5>
Explanation: <detailed judgment of prediction for "Output of Image 1" and "Output of Image 2">

Output for Answer B:
Score: <an integer score of quality from 1-5>
Explanation: <detailed judgment of prediction for "Output of Image 1" and "Output of Image 2">

Figure 8: LVLM-as-a-judge prompt in our model, which enables the model to provide the aesthetic score, pick score and alignment score for each candidate. All scores are based on the rating system, where inputs are assigned scores corresponding to their ratings. The model is fine-tuned in advance with evaluation data to understand the aesthetics of images.