# AK4Prompts: Aesthetics-driven Automatically Keywords-Ranking for Prompts in Text-To-Image Models

**Haiyang Zhang**[†] , **Mengchao Wang**[†] , **Shuai He**[†] , **Anlong Ming**[*]
School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications
{zhhy, wangmengchao, hs19951021, mal}@bupt.edu.cn

## Abstract

Current text-to-image synthesis (TIS) models have demonstrated the ability to generate high-fidelity images based on textual prompts. However, the efficacy of these models heavily relies on the keywords present in the prompts, and there is a dearth of objective analysis regarding how different keywords impact the ultimate quality of generated results. Therefore, manual evaluation becomes necessary but limited and inefficient to ascertain the role played by keywords. In this paper, we propose automated keywords-ranking for prompts (*AK4Prompts*), a keyword evaluation model based on mainstream TIS models that explicitly quantifies the multidimensional impact of various keywords on image generation based on prompts. To enable personalized keyword evaluation based on prompt content, we propose decoupling the latent representations of keywords and prompts in TIS models, followed by integrating the semantic features of prompts into keywords. For quantitative and multidimensional evaluation, we align the fused features of keywords using HPSv2, aesthetic score, and CLIP score, each representing distinct factors contributing to keyword impact. Our *AK4Prompts* can flexibly and automatically select the keywords that best match the original prompt based on individual user preferences. Extensive experimental results show the superiority of *AK4Prompts* to improve the quality of generated images significantly over strong baselines. Our approach not only enhances usability and user experience but also addresses the current gap in automated analysis and evaluation of keyword effects. Our code is available at https://github.com/mRobotit/AK4Prompts.

## 1 Introduction

Text-to-Image Synthesis (TIS) is a highly advanced and extensively utilized technique in generative Artificial Intelligence, aimed at generating realistic images based on textual input [Zhang *et al.*, 2023]. Recently, with the advancement in the modeling capabilities of large models, TIS is undergoing a revolution. Cutting-edge text-to-image diffusion models like DALLE [Ramesh *et al.*, 2021] and Latent Diffusion Models (LDMs), such as Stable Diffusion [Rombach *et al.*, 2022], have emerged as pioneers in image generation, leveraging training data at a web scale. Additionally, Consistency models [Song *et al.*, 2023], a new class of generative models capable of producing high-quality samples through a single network evaluation, have inspired further advancements. Building upon this concept, Luo et al. [Luo *et al.*, 2023a] propose Latent Consistency Models (LCMs), which enable rapid inference with minimal steps on any pre-trained LDM and significantly enhance the speed of TIS. However, the image generation quality in these existing models relies heavily on the sophisticated design of keyword-based text prompts [Zhong *et al.*, 2023]. This reliance stems from the training data quality, necessitating detailed prompts to produce high-quality images [Betker *et al.*, 2023]. Determining the most effective keywords for a given prompt content often requires iterative experimentation and research within online communities [Oppenlaender, 2022]. In real-world scenarios, individuals lacking expertise often encounter challenges in selecting appropriate keywords for composing detailed prompts. This necessity leads to subjective assessments of the impact and quality of chosen keywords through repeated generation attempts, resulting in significant time and resource losses.

Prompt engineering is an emerging research field focused on developing more effective prompts for deep generative models. In the TIS community, a common practice involves investigating the effects of various keywords. BestPrompt [Pavlichenko and Ustalov, 2023] initially explores these effects by employing human evaluation to assess a limited number of text prompts and keywords, ultimately selecting a set of exceptional keywords. However, this approach has inherent limitations and inefficiencies, posing challenges in comprehensively and objectively exploring the impact of all keywords on generated results. Furthermore, image generation is jointly controlled by the interaction between keywords and basic prompt content. However, this study only considers the individual influence of keywords, neglecting potential variations in their effects across different prompt categories. Another prevalent approach involves fine-tuning a language model for TIS prompt generation using com-

[*]Corresponding author.

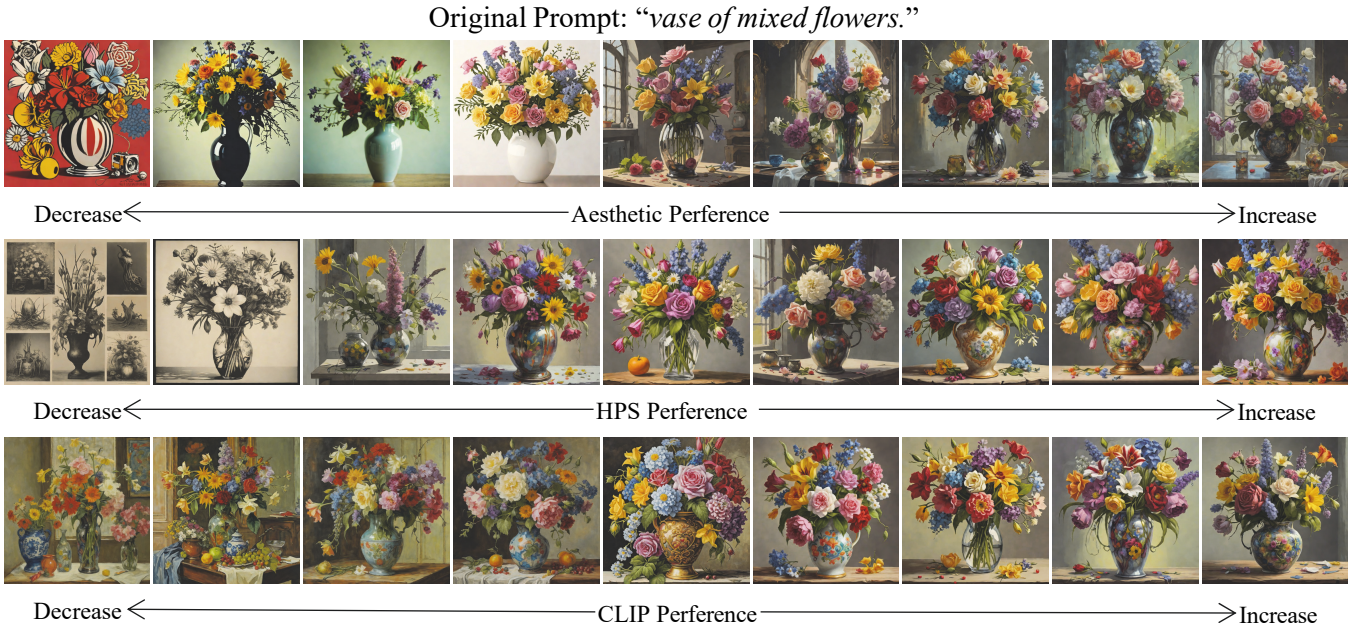


Figure 1: The evolution of images is facilitated by customizing keyword preferences with *AK4Prompts*, leveraging the SDXL-Turbo model.

plex prompts datasets, such as MagicPrompt [1] and BeautifulPrompt [Cao *et al.*, 2023]. However, these methods primarily focus on completing or rewriting prompts, potentially altering the content entities. Moreover, the generated prompts often lack flexibility and exhibit excessive rigidity.

In this paper, we propose a keyword evaluation model named *AK4Prompts* that can explicitly quantify the multidimensional impact of different keywords on the quality of generated images in response to user prompts. Our *AK4Prompts* automatically selects the keywords that best match the original prompt based on users' preferences, effectively enhancing the quality of the generated images. Our method preserves the content of the original prompt and allows users to customize their final outputs, thus overcoming the limitations of previous approaches. Meanwhile, it significantly alleviates the complexity of prompt design and the difficulty of selecting keywords, thereby saving users considerable trial and error time.

Figure 1 illustrates the evolution of images generated by setting different degrees of specific preferences. In summary, the main contributions of this study are as follows:

- We curate a high-quality set of keywords and develop an evaluation framework based on established TIS models. This framework enables automated and efficient assessment of these keywords, eliminating the need for manual intervention.
- We propose *AK4Prompts*, a keyword evaluation model that can explicitly quantify the multidimensional impact of different keywords on the quality of images generated by TIS models based on user prompts.
- Our approach demonstrates the judicious selection of the most appropriate keywords based on user prompts and preferences. Extensive experimental results showcase its superiority over strong baselines.

## 2 Related Work

### 2.1 Text-to-Image Synthesis (TIS)

TIS is a multi-modal task involving the generation of images based on textual inputs. In the early years, popular image generation tasks were mainly based on Generative Adversarial Network (GAN) [Goodfellow *et al.*, 2014; Reed *et al.*, 2016]. More recently, diffusion models [Ho *et al.*, 2020; Song *et al.*, 2020; Sohl-Dickstein *et al.*, 2015; Dhariwal and Nichol, 2021] have achieved remarkable results. Subsequently, LDMs (e.g., Stable Diffusion [Rombach *et al.*, 2022]) apply a diffusion model in the latent space using a pre-trained autoencoder for efficient training and inference. This development has resulted in the emergence of high-performance generative models trained on billions of data points, including DALLE-2 [Ramesh *et al.*, 2022], Imagen [Saharia *et al.*, 2022], Imagic [Kawar *et al.*, 2023], and SDXL [Podell *et al.*, 2023]. Among these models, the Stable Diffusion series stands out as open-source models with an active user community.

Recently, LCMs [Luo *et al.*, 2023a], LCM-LoRA [Luo *et al.*, 2023b], and SDXL-Turbo [Sauer *et al.*, 2023] have emerged as solutions to the slow sampling issue in image generation, inspired by Consistency Models [Song *et al.*, 2023]. LCM-LoRA serves as a universal, training-free acceleration module. It can be directly integrated into various Stable Diffusion fine-tuned models or LoRAs [Hu *et al.*, 2021], facilitating fast inference with minimal steps. This allows us to efficiently infer and explore a large number of keyword and

[1] https://huggingface.co/Gustavosta/MagicPrompt-Stable-Diffusion

prompt category combinations.

However, the qualities of generated images are dependent on the keywords provided in the prompts. TIS models usually utilize a text encoder, which employs pre-trained language models like CLIP [Radford *et al.*, 2021] to convert textual inputs into latent vectors. These latent vectors associated with individual keywords represent specific control influences on images, which can be utilized for our quantitative evaluation task.

### 2.2 TIS Evaluation

The aesthetic score [Schuhmann *et al.*, 2022] assesses the aesthetic quality of individual images, while the CLIP score [Radford *et al.*, 2021] quantifies the similarity between generated images and prompts. Additionally, there exist metrics specifically trained to align with human preferences, such as [Wu *et al.*, 2023b; He *et al.*, 2022; Wu *et al.*, 2023a; Xu *et al.*, 2023; Kirstain *et al.*, 2023; He *et al.*, 2023a; He *et al.*, 2023b]. Human preferences are intricate and encompass various dimensions, such as text-image similarity, image fidelity, aesthetics, and other factors. Among these metrics, HPSv2 stands out for its consistent scoring and use of larger, more diverse training datasets. Significantly, the CLIP model has been fine-tuned on an extensive dataset, featuring 798,090 human ranking choices across 433,760 image pairs [Wu *et al.*, 2023a]. These evaluation metrics provide valuable visual feedback and enable a quantitative assessment of the multidimensional impact of keywords on TIS.

In this paper, we adopt a quantitative and multidimensional approach to evaluate the fused features of keywords. We use multiple TIS evaluation metrics, including HPSv2, the aesthetic score, and the CLIP score, with each metric representing distinct aspects of keyword effects.

### 2.3 Prompt Engineering for TIS

Due to the remarkable potential of TIS, there has been a surge of interest in prompt engineering. MagicPrompt, a widely adopted model for automatic prompt completion, is trained on high-quality prompts sourced from the Internet. BeautifulPrompt [Cao *et al.*, 2023] utilizes a reinforcement learning technique, complemented by visual AI feedback, to fine-tune language models and maximize the rewards of generated prompts. However, these methods primarily focus on completing or rewriting prompts, which may inadvertently alter content entities, resulting in rigid and inflexible generated prompts. Liu *et al.* [Liu and Chilton, 2022] conducted a series of experiments, proposing several design guidelines for text-to-image prompt engineering that confirm the beneficial impact of incorporating keywords. Additionally, [Oppenlaender, 2023] identified six distinct types of keywords through an ethnographic study spanning three months within the online generative art community.

However, previous studies have been constrained by the laborious and time-consuming manual engineering of prompts. In contrast, BestPrompt [Pavlichenko and Ustalov, 2023] employs a genetic algorithm to identify keywords that compose prompts, aiming to achieve optimal aesthetic quality in images. Nevertheless, this study solely relies on selecting a fixed set of keywords based on average performance determined through human evaluation.

Conversely, *AK4Prompts* can assess the varying impact of keywords and select the most suitable ones for specific prompts, thereby enhancing the visual appeal of generated images according to user preferences. To address the potential variations in the effects of keywords across different prompt categories and provide personalized keyword evaluation based on prompt content, we decouple the latent representations of keywords and prompts in TIS models, which have distinct impacts on the generated images. Subsequently, to further enhance this integration, we employ multiple cross-attention layers to merge the semantic features of prompts with those of keywords.

## 3 Data Collection for Training

**Simple Prompt datasets.** The image-caption datasets, such as COCO Captions [Chen *et al.*, 2015] and LAION [Schuhmann *et al.*, 2022] primarily describing real images, may not align with the interests of generative model users. DiffusionDB [Wang *et al.*, 2022] comprises a large-scale dataset filled with a wide range of user-written prompts and generated images. However, it features numerous complex keyword-based text prompts. Regarding our prompt dataset, we require simple text prompts to maximize the impact of different keywords. SUR-adapter [Zhong *et al.*, 2023] employs a pre-trained BLIP model to generate captions for images associated with high-quality prompts, treating the brief and less detailed resulting captions as low-quality prompts. BeautifulPrompt [Cao *et al.*, 2023] enhances this approach by using ChatGPT [OpenAI, 2023] to summarize high-quality prompts, treating these summaries as low-quality prompts. The dataset includes 143k simple prompts and 2k test prompts. We train our model using BeautifulPrompt's training set and evaluate its performance on the test set.

**Keywords Collection.** We extract a series of keywords from complex prompts in DiffusionDB, separating them by commas and sorting them based on frequency of occurrence. After manually removing non-keywords and those with lower frequencies, we obtain a list of 51,291 keywords. However, the keyword set contains numerous semantically consistent and repetitive terms. Since the SD utilizes CLIP to encode text inputs into latent vectors for conditional control, keywords with similar CLIP encodings have comparable effects on the generation outcome. To reduce keyword redundancy and prevent the model from developing a specific bias, we employ CLIP to uniformly encode the keywords, compute their similarity matrix, and apply the k-means clustering algorithm for grouping them into clusters. We retain the most frequently occurring keywords in each cluster while removing redundant ones to reduce the overall number of keywords further. Through extensive experimentation, we ultimately select 954 distinct semantic representations as our final keyword set. Our *AK4Prompts* can generalize to all 51,291 keywords via CLIP encoding. Keywords with similar semantics receive similar scores, even rare ones, matched based on learned semantic similarities.

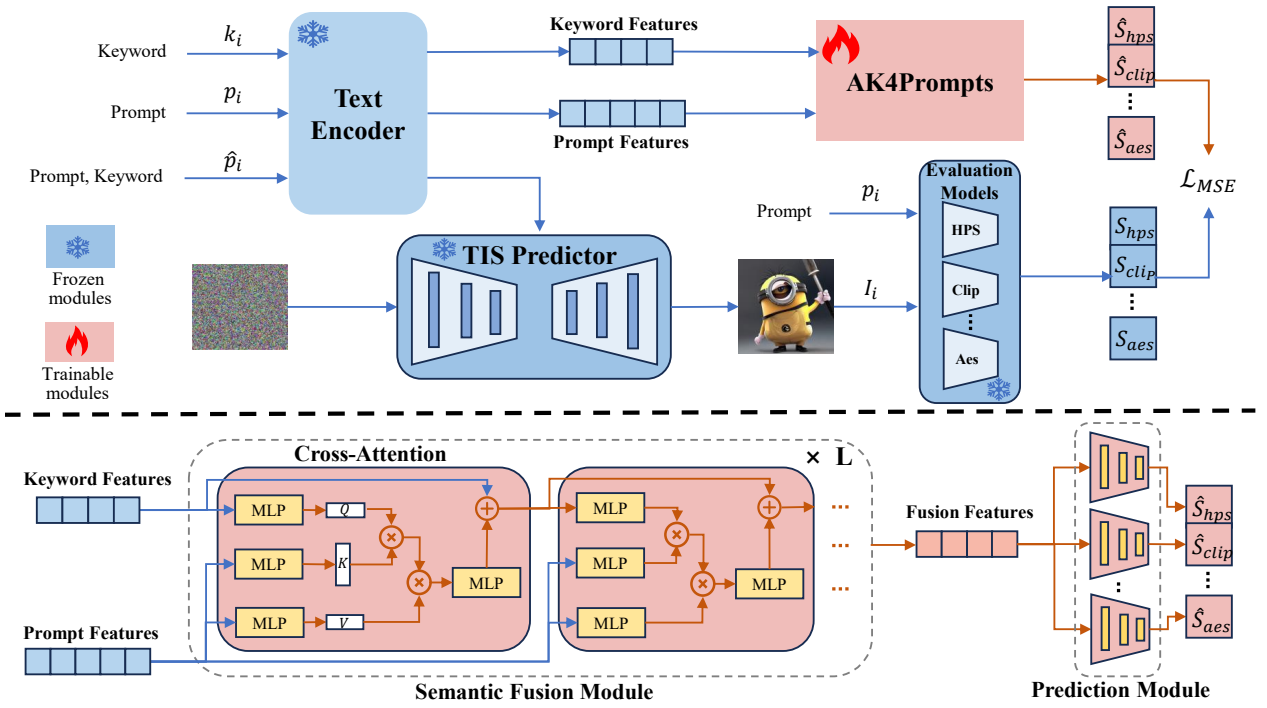


Figure 2: Illustration of *AK4Prompts*. (Top) The overall architecture of our training pipeline, where only the newly added module (highlighted in red) is trained while the pretrained text-to-image model remains frozen. (Bottom) The network structure of our module, which utilizes the *AK4Prompts* module to assess the impact of keywords on image quality generated in response to user prompts.

## 4 AK4Prompts

As shown in Figure 2 (Top), let us consider the keyword-prompt pairs $(p_i, k_i)_{i=1}^{N}$, where $p_i$ represents the original prompt and $k_i$ denotes a randomly selected keyword from the collected set of keywords. Firstly, we freeze all learnable parameters of the text encoder $f_{En}$ and the predictor $f_{pre}$ in the pre-trained TIS model, as well as all the evaluation models used. Then, we utilize the TIS model to generate images $I_i$ corresponding to keyword-based prompt $\hat{p}_i$ and calculate average scores for both images $I_i$ and the original prompt $p_i$ using various evaluation models.

Finally, we propose our trainable AK4Prompts model to predict these scores based on the features of $p_i$ and $k_i$ from the text encoder. Our model, as illustrated in Figure 2 (Bottom), comprises two components: the Semantic Fusion Module (Section 4.1) and the Multidimensional Score Prediction Module (Section 4.2). Section 4.3 demonstrates how to utilize AK4Prompts for inference by identifying suitable keywords according to user prompts and customized preferences.

### 4.1 Semantic Fusion of Keywords and Prompts

The structure of the Semantic Fusion Module is depicted in Figure 2 (Bottom), which comprises $L$ cross-attention layers. These layers are responsible for fusing the semantic features of keywords and prompts, generating attention maps for each textual token of prompts. Each layer consists of four learnable transformations, denoted as $h_j(\cdot)$ for $j = 1, 2, 3, 4$, implemented using a multi-layer perceptron (MLP). We utilize the outputs $f_{En}(k_i)$ and $f_{En}(p_i)$ from the text encoder as initial embeddings for keyword features $emb_0^{k_i}$ and prompt features $emb^{p_i}$, respectively.

$$emb_0^{k_i} = f_{En}(k_i), emb^{p_i} = f_{En}(p_i). \quad (1)$$

Then, we construct $Q^i = h_3[emb_\ell^{k_i}]$ and $K^i = h_2[emb^{p_i}]$, and calculate attention values as described in [Devlin *et al.*, 2018; Vaswani *et al.*, 2017]

$$Q_\ell^i = h_3(emb_{\ell-1}^{k_i}), K_\ell^i = h_2(emb^{p_i}) \qquad \ell = 1...L, \quad (2)$$

$$\text{att}_\ell^i = \text{softmax}(\frac{Q_\ell^i {K_\ell^i}^T}{\sqrt{d}}) \qquad \ell = 1...L, \quad (3)$$

where $d$ represents the feature dimension of $Q_i$ and $K_i$, and the cell $\text{att}_\ell^{i,j}$ defines the weight of the value of the j-th token in $p_i$ on $k_i$. Additionally, we construct $V_i = h_1[emb^{p_i}]$, and obtain the calibrated semantic information as $\hat{V}_i = V_i \otimes \text{att}_i$, which is then used to update the keyword features $emb^{k_i}$. Afterwards, the output of the first cross-attention sublayer is transformed by a learnable transformation function denoted as $h_4(\cdot)$ to obtain the fused semantic features.

$$V_\ell^i = h_1(emb^{p_i}), \hat{V}_\ell^i = V_\ell^i \otimes \text{att}_\ell^i \qquad \ell = 1...L, \quad (4)$$

$$emb_\ell^{k_i} = LN(emb_{\ell-1}^{k_i} + h_4(\hat{V}_\ell^i)) \qquad \ell = 1...L. \quad (5)$$

Here, We employ a residual connection [He *et al.*, 2016] around each of the two sub-layers, followed by layer normalization [Ba *et al.*, 2016]. Finally, the output of the Semantic Fusion Module is obtained as $emb_L^{k_i}$ through the cross-attention layers.

| TIS Model | Method | Aesthetic Score | | | HPSv2 | | | CLIP Score | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Top-1 | Top-5 | Top-10 | Top-1 | Top-5 | Top-10 | Top-1 | Top-5 | Top-10 |
| SD (1.5) + LCM-LoRA | Original | 5.70 | 5.70 | 5.70 | 0.256 | 0.256 | 0.256 | 0.258 | **0.258** | **0.258** |
| | BestPrompt | 5.80 | 5.97 | 6.12 | 0.257 | 0.256 | 0.260 | 0.255 | 0.245 | 0.243 |
| | MostPopular | 5.76 | 6.04 | 6.19 | 0.256 | 0.257 | 0.257 | 0.259 | 0.251 | 0.246 |
| | *AK4Prompts (ours)* | **5.82** | **6.14** | **6.23** | **0.258** | **0.261** | **0.261** | **0.259** | 0.252 | 0.247 |
| SD (1.5) | Original | 5.70 | 5.70 | 5.70 | 0.261 | 0.261 | 0.261 | **0.287** | **0.287** | **0.287** |
| | BestPrompt | 5.80 | 6.14 | 6.26 | 0.261 | 0.263 | 0.265 | 0.283 | 0.276 | 0.270 |
| | MostPopular | 5.80 | 6.19 | 6.37 | 0.261 | 0.261 | 0.260 | 0.286 | 0.277 | 0.272 |
| | *AK4Prompts (ours)* | **6.00** | **6.29** | **6.43** | **0.262** | **0.263** | **0.266** | 0.286 | 0.278 | 0.270 |
| SDXL-Turbo | Original | 6.21 | 6.21 | 6.21 | 0.275 | 0.275 | 0.275 | **0.286** | **0.286** | **0.286** |
| | BestPrompt | 6.27 | 6.46 | 6.42 | 0.275 | 0.276 | 0.277 | 0.278 | 0.273 | 0.268 |
| | MostPopular | 6.31 | 6.53 | 6.62 | 0.275 | 0.275 | 0.276 | 0.282 | 0.272 | 0.269 |
| | *AK4Prompts (ours)* | **6.45** | **6.57** | **6.71** | **0.276** | **0.276** | **0.278** | 0.282 | 0.275 | 0.276 |

Table 1: Results on the testing set. "Original" refers to the method that directly sends the original prompts to TIS models without modification. "Top-n" refers to using the top-n keywords from different methods.

### 4.2 Multidimensional Score Prediction Module

The Score Prediction Module consists of three MLP score prediction heads, $\hat{S}_{hps}(\cdot)$, $\hat{S}_{clip}(\cdot)$ and $\hat{S}_{aes}(\cdot)$, as shown in Figure 2 (Bottom). Each prediction head utilizes the output keyword embedding from the Semantic Fusion Module to predict the average HPSv2 score [Wu *et al.*, 2023a], CLIP similarity score [Radford *et al.*, 2021], and aesthetic score [Schuhmann *et al.*, 2022] of the generated image based on the corresponding keyword prompt.

Briefly, to enhance the prompt $p_i$ with keyword $k_i$, append $k_i$ to the end of $p_i$ and form $\hat{p_i}$. Then input $\hat{p_i}$ into the downstream TIS model to generate image $I_i$:

$$I_i = f_{pre}(f_{En}(\hat{p_i}), seed). \tag{6}$$

HPSv2 is a human preference model trained on a large dataset of text-to-image prompts and real user preferences. To mitigate the impact of random seeds on the quality of images generated by the TIS model, we employ 8 different random seeds to generate images and average the results. The calculated averaged HPSv2 score $HPS$ is used as the ground truth for training our model. The loss function is as follows:

$$\mathcal{L}_{hps} = -\frac{1}{N}\sum_{i}^{N} MSE(\hat{S}_{hps}(emb_L^{k_i}), HPS), \tag{7}$$

where $\hat{S}_{hps}(emb_L^{k_i})$ represents the scalar output of the HPSv2 prediction head for the keyword embedding $emb_L^{k_i}$ obtained from the Semantic Fusion Module. MSE refers to Mean Squared Error, and $N$ denotes the total number of samples.

CLIP score is a metric used to evaluate the correlation between the image generated by the model and the original text. In our work, we use a MLP prediction head to assess the consistency between the original prompt $p_i$ and generated image $I_i$ under the influence of keyword $k_i$. The calculated average CLIP score $CLIP$ is used as the ground truth to train our model. The loss function is:

$$\mathcal{L}_{clip} = -\frac{1}{N}\sum_{i}^{N} MSE(\hat{S}_{clip}(emb_L^{k_i}), CLIP). \tag{8}$$

The aesthetic score model is trained to predict the rating that people give when asked "how much do you like this image on a scale from 1 to 10". To maintain consistency with $HPS$ and $CLIP$, we scale the aesthetic scores to a range of 0 to 1. Similarly, a MLP prediction head is trained to fit the corresponding keywords to the aesthetic scores $AES$ of the generated images:

$$\mathcal{L}_{aes} = -\frac{1}{N}\sum_{i}^{N} MSE(\hat{S}_{aes}(emb_L^{k_i}), AES). \tag{9}$$

Finally, we combine the losses of the three MLP prediction heads as the final $\mathcal{L}_{total}$ using loss coefficients $h$, $a$, and $c$.

$$\mathcal{L}_{total} = h \cdot \mathcal{L}_{hps} + c \cdot \mathcal{L}_{clip} + a \cdot \mathcal{L}_{aes}. \tag{10}$$

### 4.3 Customized Keywords-Ranking

After training, our *AK4Prompts* $g_{ak}(\cdot)$ can predict scores for each keyword based on simple prompts, allowing us to determine the optimal set of keywords that align with these prompts. For a given simple prompt $p$, we input it into our model to simultaneously predict HPSv2 scores, CLIP scores, and aesthetic scores for all keywords in a parallel inference step:

$$S_{hps}^i, S_{clip}^i, S_{aes}^i = g_{bk}(p, k^i; \phi_1, \phi_2) \quad i = 1...K, \tag{11}$$

where $k^i$ represents a keyword and $K$ denotes the total number of keywords. Then, we can calculate a comprehensive score for each keyword $k^i$ as follows:

$$S_{avg}^i = \hat{h} * S_{hps}^i + \hat{c} * S_{clip}^i + \hat{a} * S_{aes}^i \quad i = 1...K, \tag{12}$$

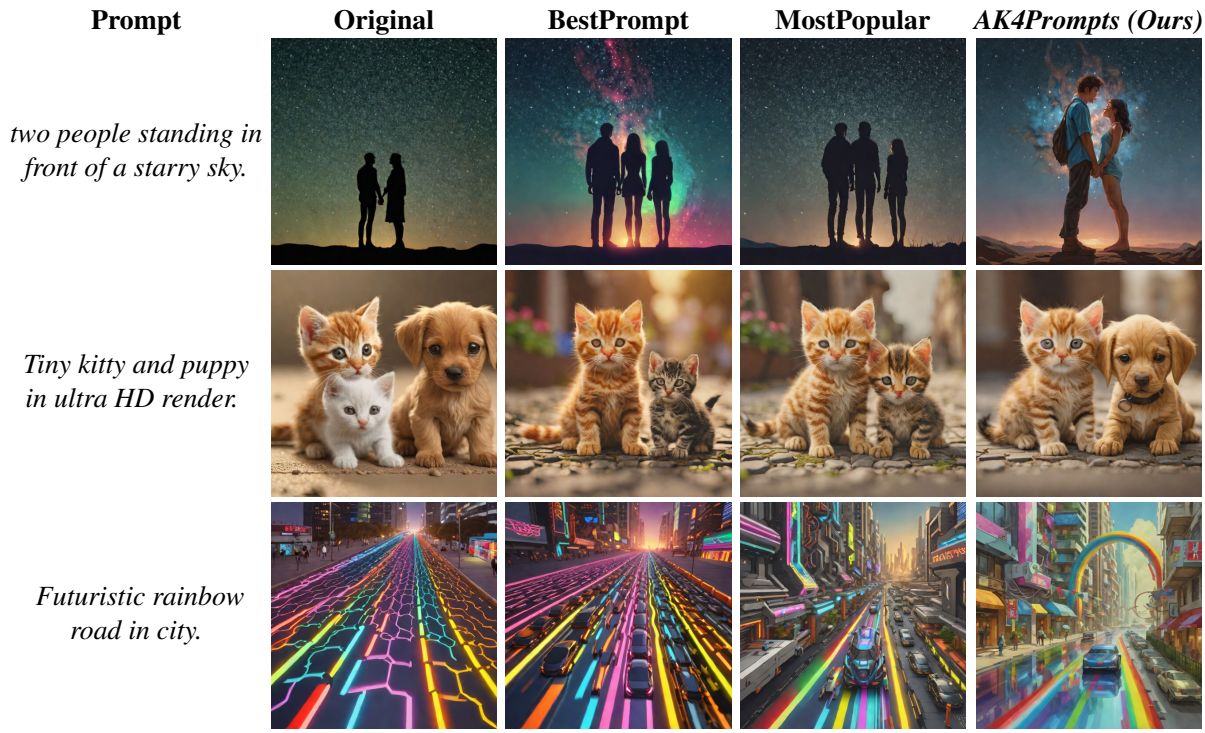


Figure 3: The qualities of images generated from prompts with different keyword selection methods using the SDXL-Turbo model.

where $(\hat{h}, \hat{c}, \hat{a})$ represent the weights for the HPS score, CLIP score, and aesthetic score, respectively.

Finally, a comprehensive ranking of all keywords can be obtained, where higher-ranked keywords exhibit a more pronounced enhancing effect on the prompt $p$. Moreover, by customizing the values of $(\hat{h}, \hat{c}, \hat{a})$ based on user preferences, keywords with different emphases can be sorted and selected. For instance, assigning a larger value to $\hat{a}$ enables the selection of keywords that exert a greater impact on aesthetics. Based on this keyword ranking, we can choose the top-performing keywords and append them to simple prompts in order to enhance overall image generation performance.

Our method achieves higher computational efficiency compared to prompt generation models due to its use of precomputed semantic embeddings, avoidance of self-attention computation during keyword processing, and smaller model size.

## 5 Experiments

**Training Settings.** Our model was trained on 143K BeautifulPrompt training samples. We utilized the pre-trained stable diffusion 1.5 as our TIS model and employed LCM-LoRA for accelerated inference. We generated 512x512 resolution images through a four-step inference with a CFG scale $\omega$ set to 1.0, leveraging FLOAT16 formats to save GPU memory and speed up training. For the semantic fusion module, we set $L = 3$. Regarding $\mathcal{L}_{total}$, we set $h = 2.25$, $c = 2.25$, and $a = 1$. Due to different ranges in the distributions of these scores, we assigned a larger weight to the one with smaller ranges to balance gradients and enhance prediction accuracy. The learning rate was set at 1e-4, weight decay at 1e-2, batch size at 32, and training step at 88,000 steps. All experiments were implemented in PyTorch and run on a single server with NVIDIA RTX3090TI GPUs.

**Baselines.** We consider two strong baselines: BestPrompt [Pavlichenko and Ustalov, 2023] and MostPopular. BestPrompt is the first work to identify the best keywords for TIS, using a genetic algorithm that identifies a set of high-quality keywords through extensive human evaluations. MostPopular represents the most frequently used keywords in TIS, which have gained widespread recognition within the user community for their proven effectiveness in enhancing the quality of generated images. Our method requires setting the number of selected keywords and specific preferences. Whereas the generation methods may produce new prompts with different keyword lengths, making comparison challenging.

**Evaluation Protocols.** We assess the images generated by the TIS model using 2K BeautifulPrompt testing prompts employing various keyword selection methods. Multiple TIS models are employed to generate images, and we compute the aesthetic score [Schuhmann *et al.*, 2022], HPSv2 [Wu *et al.*, 2023a], and CLIP score [Radford *et al.*, 2021] for both the images and original prompts. For the baselines, we evaluate using different numbers of keywords, specifically top-1, top-5, and top-10. Regarding our *AK4Prompts* approach, we set ($\hat{h}$=3, $\hat{c}$=5, $\hat{a}$=1) to determine a comprehensive ranking for all keywords and select an equivalent number of top keywords for evaluation and comparison purposes. As a comparative measure due to baselines' emphasis on clip scores, we also assign a higher weight to $\hat{c}$ in order to enhance the clip score.

| TIS Model | Method | Aesthetic Score | HPSv2 | CLIP Score |
|---|---|---|---|---|
| SD (1.5)+ LCM-LoRA | Original | 5.70 | 0.256 | **0.258** |
| | BestPrompt | 5.97 | 0.256 | 0.245 |
| | MostCommon | 6.04 | 0.257 | 0.251 |
| | *AK4Prompts* | **6.49** | **0.261** | 0.254 |
| SD (1.5) | Original | 5.70 | 0.261 | **0.287** |
| | BestPrompt | 6.14 | 0.263 | 0.276 |
| | MostCommon | 6.19 | 0.261 | 0.277 |
| | *AK4Prompts* | **6.66** | **0.265** | 0.283 |
| SDXL-Turbo | Original | 6.21 | 0.275 | **0.286** |
| | BestPrompt | 6.46 | 0.277 | 0.273 |
| | MostCommon | 6.53 | 0.275 | 0.272 |
| | *AK4Prompts* | **7.03** | **0.279** | 0.283 |

Table 2: Performance of selecting the most impactful keywords on the aesthetic score, CLIP score, and HPSv2 score, respectively. Each score is calculated independently based on the top-5 highest-scoring keywords.

Additionally, we set ($\hat{h}$, $\hat{c}$, $\hat{a}$) as (1, 0, 0), (0, 1, 0), and (0, 0, 1) respectively to select the top-5 keywords that have a significant impact on the aesthetic score, CLIP similarity score, and human preference score for evaluation.

## 5.1 Overall Results

Our method consistently outperforms the other baselines in most scores, as demonstrated by Table 1. Given that the CLIP score reflects semantic consistency between text and image, it is unsurprising that sending the original prompts without any additional keywords to Stable Diffusion unchanged yields the highest score. Our approach achieves superior aesthetic scores, HPSv2, and CLIP scores compared to the baselines while maintaining a comparable CLIP score to the Original prompt. This highlights the effectiveness of our proposed method, *AK4Prompts*, in selecting optimal keywords for input prompts.

Our *AK4Prompts* is applicable to all SD-style models. The experimental results of SD1.5 and SDXL-Turbo demonstrate the transferability of *AK4Prompts* to other diffusion-based TIS models. Taking into consideration the widely adopted model SDXL-Turbo [2], Figure 3 illustrates that, despite already performing well with most vanilla prompts, SDXL-Turbo can further enhance its image generation capabilities when utilizing *AK4Prompts*. As depicted in Table 2, the selection of only the top 5 keywords, predicted by the model with the highest aesthetic scores for image generation, leads to a significant enhancement in aesthetic scores, as well as CLIP and HPS scores. This observation signifies that our approach proficiently captures diverse keyword effects, offering guidance to users and enabling customization of final output preferences. Additional examples are provided in Appendix A and B in our code repository, including qualitative analyses of different keywords and the specific effects of different categories of keywords on image generation. We also feature a qualitative comparison of images generated using keywords selected with different preferences.

[2]https://huggingface.co/stabilityai/sdxl-turbo

## 5.2 Ablation Study

In Table 3, we analyze the various design choices implemented throughout our experiments. As discussed in Section 4.1, the Semantic Fusion Module integrates the semantics of simple prompts with keywords. This integration enables a more effective combination of different prompt semantics, thereby providing personalized keywords. To validate the module's effectiveness, we initially compare it with a simpler method that omits the Semantic Fusion Module. This method involves directly using the text encoder to obtain semantic encodings of user prompts and additional keywords, followed by score prediction using MLP prediction heads. Subsequently, we employ self-attention instead of cross-attention for integrating semantic features from user prompts and additional keywords. Finally, we evaluate different scheduling steps using LCM-LoRA. For a fair comparison, both models were trained for 88,000 steps under identical configurations.

| | Aesthetic Score |
|---|---|
| *AK4Prompts (ours)* | **6.49** |
| w/o SF | 6.34 |
| with self-attention | 6.43 |
| with step=1 | 6.28 |

Table 3: The proposed method consistently outperforms all other baselines in terms of achieved rewards, as evidenced by the table featuring the top-5 keywords.

# 6 Conclusion

In this paper, we propose a universal algorithmic framework named *AK4Prompts* that quantitatively assesses the different effects of various keywords in TIS models based on simple prompt. Our approach significantly reduces the complexity of prompt design and the difficulty of keyword selection, thereby saving users considerable trial-and-error time. While *AK4Prompts* can automatically select keywords for user prompts and generate more aesthetically pleasing images, we find it challenging to strike a balance among factors such as image aesthetics, human preferences, and text-image consistency in order to choose the best keywords. At times, in pursuit of better text-image consistency, we have to compromise on certain aspects of image aesthetics. Additionally, since we use SD1.5 and employ LCM-LoRA with CFG scale $\omega$ set to 1.0, only positive keywords are used to control generation while negative keywords are not considered. Similar to positive prompts, negative prompts are also crucial inputs for many TIS models. The final image generation results depend on both positive and negative prompts. Therefore, discovering the best negative keywords can effectively improve the quality of generated images as well. The incorporation of negative keywords into our framework is achievable, but addressing this aspect will be a key focus in the development of AK4Prompts 2.0.


# Acknowledgements

This work was supported by the Innovation Research Group Project of NSFC (61921003).

## Contributions

† on Page 1 indicates that Haiyang Zhang, Mengchao Wang and Shuai He, three authors of this paper, have made equal contributions.

## References


[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Betker *et al.*, 2023] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023.

[Cao *et al.*, 2023] Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. *arXiv preprint arXiv:2311.06752*, 2023.

[Chen *et al.*, 2015] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2022] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, pages 942–948, 2022.

[He *et al.*, 2023a] Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21838–21847, 2023.

[He *et al.*, 2023b] Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma. Eat: An enhancer for aesthetics-oriented transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1023–1032, 2023.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Kawar *et al.*, 2023] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.

[Kirstain *et al.*, 2023] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023.

[Liu and Chilton, 2022] Vivian Liu and Lydia B Chilton. 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, 2022.

[Luo *et al.*, 2023a] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[Luo *et al.*, 2023b] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

[OpenAI, 2023] OpenAI. Chatgpt. *https://openai.com/chatgpt*, 2023.

[Oppenlaender, 2022] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, Academic Mindtrek '22, page 192–202, New York, NY, USA, 2022. Association for Computing Machinery.

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

[Pavlichenko and Ustalov, 2023] Nikita Pavlichenko and Dmitry Ustalov. 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, 2023.

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

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

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

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

[Sauer *et al.*, 2023] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

[Schuhmann *et al.*, 2022] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Song *et al.*, 2023] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2022] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

[Wu *et al.*, 2023a] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[Wu *et al.*, 2023b] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 2023.

[Xu *et al.*, 2023] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

[Zhang *et al.*, 2023] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[Zhong *et al.*, 2023] Shanshan Zhong, Zhongzhan Huang, Weushao Wen, Jinghui Qin, and Liang Lin. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 567–578, 2023.