Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis

Xiaoshi Wu[1*], **Yiming Hao**[5*], **Keqiang Sun**[1], **Yixiong Chen**[2],
Feng Zhu[3], **Rui Zhao**[3,4], **Hongsheng Li**[1,5]

[1]Multimedia Laboratory, The Chinese University of Hong Kong [2]CUHK-SZ, SRIBD
[3]SenseTime Research [4]Qing Yuan Research Institute, Shanghai Jiao Tong University
[5]Centre for Perceptual and Interactive Intelligence (CPII)
{wuxiaoshi@link, kqsun@link, hsli@ee}.cuhk.edu.hk
ymhao@cpii.hk, yixiongchen@link.cuhk.edu.cn
{zhufeng, zhaorui}@sensetime.com

Abstract

Recent text-to-image generative models can generate high-fidelity images from text inputs, but the quality of these generated images cannot be accurately evaluated by existing evaluation metrics. To address this issue, we introduce Human Preference Dataset v2 (HPD v2), a large-scale dataset that captures human preferences on images from a wide range of sources. HPD v2 comprises 798,090 human preference choices on 430,060 pairs of images, making it the largest dataset of its kind. The text prompts and images are deliberately collected to eliminate potential bias, which is a common issue in previous datasets. By fine-tuning CLIP on HPD v2, we obtain Human Preference Score v2 (HPS v2), a scoring model that can more accurately predict text-generated images' human preferences. Our experiments demonstrate that HPS v2 generalizes better than previous metrics across various image distributions and is responsive to algorithmic improvements of text-to-image generative models, making it a preferable evaluation metric for these models. We also investigate the design of the evaluation prompts for text-to-image generative models, to make the evaluation stable, fair and easy-to-use. Finally, we establish a benchmark for text-to-image generative models using HPS v2, which includes a set of recent text-to-image models from the academia, community and industry. The code and dataset is / will be available at https://github.com/tgxs002/HPSv2

1 Introduction

Recent advancements in text-to-image synthesis [16, 15, 17, 12, 2] have made it possible to generate high-fidelity images based on text prompts. However, images generated with different random seed often have varying quality, and previous works [24, 25, 7] demonstrate that popular metrics, such as Inception Score (IS) [19], Fréchet Inception Distance (FID) [5], and CLIP Score [14], do not correlate well with human preferences on these images. Therefore, human preference is an important but poorly tracked aspect of text-to-image generative models. To facilitate research in this area, we construct a large-scale dataset annotated with human preferences, namely Human Preference Dataset v2 (HPD v2). We also established a benchmark based on a preference prediction model, Human Preference Score v2 (HPS v2), which is trained on HPD v2, to validate algorithmic developments for human-aligned image synthesis.

[*]Equal contribution.

Preprint. Under review.

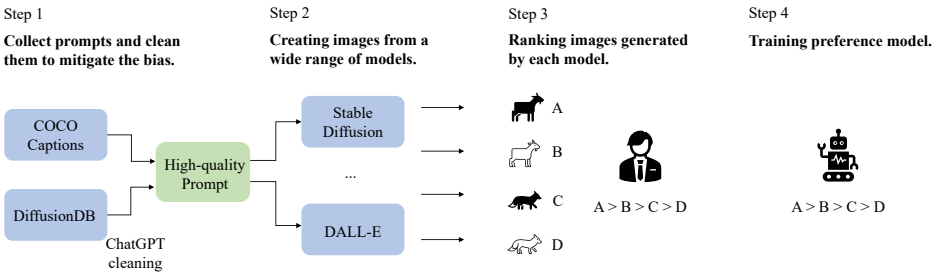


Figure 1: Overview of this work. We firstly collect Human Preference Dataset v2 (HPD v2) and then train a preference prediction model, Human Preference Score v2 (HPS v2), on it.

HPD v2 is a comprehensive dataset of human preferences for images sourced from a wide range of text-to-image generative models and the COCO Captions dataset [1]. It comprises 798k human-annotated pairwise comparisons of images generated from the same prompt, making it the largest dataset of its kind. The dataset is collected with special attention to potential bias. The first type of bias derives from the image source. Previous datasets [24, 25, 7] mainly contain images generated from Stable Diffusion [17] and its variants. Therefore, we cannot verify whether the models trained and evaluated exclusively on these datasets can generalize to other image distributions. HPD v2 incorporates images generated from 9 recent text-to-image generative models [2, 12, 3, 9, 27, 17, 4, 16], as well as real images from the COCO Captions dataset [1], which allows for the examination of biases introduced by limited image sources. Another source of bias lies in text prompts. User-written prompts, such as prompts in DiffusionDB [23], often follow a specific organization of description plus several style words, where the style words often contain contradictions, making it harder for annotators to understand. (see Tab. 1 for examples) The style words are also highly biased, leading to issues in training and evaluation. To tackle this prompt bias, we employ ChatGPT to remove style words and organize the prompt into one clearly written sentence.

Based on HPD v2, a preference prediction model, Human Preference Score v2 (HPS v2), is trained to estimate the human preference probability on images generated from text prompts. Our HPS v2 has better generalization capability than previous models, including HPS v1 [24], ImageReward [25] and PickScore [7], and it can serve as a better evaluation metric for text-to-image generative models.

HPS v2 estimates the probability of an image to be preferred by viewers against other images, and thus the probability logit can be utilized to evaluate a text-to-image generative model's ability to generate preferable images. However, HPS v2 alone is not enough for a comprehensive evaluation, as the evaluation is also influenced by testing text prompts. We therefore further study the design of the evaluation prompts, to make it fair, stable, and easy-to-use. Imagen [18] proposes to evaluate on DrawBench, which only provides a list of prompts, but there is not an automatic way to evaluate on them. DiffusionDB [23] is a large scale database of user-written prompts and images generated from them. However, a large portion of the prompts are biased towards certain dataset-specific style words, and is not fair for models trained with other data. To achieve a fair comparison, we set up a benchmark that evaluates a text-to-image generative model by its ability of generating images conditioned on less biased prompts of four general styles: "Animation", "Concept-art", "Painting", and "Photo". We carefully study the number of prompts for evaluation to ensure that the average score is statistically stable. With the help of the deliberately designed evaluation prompts, we evaluate a relatively complete list of recent text-to-image generative models to set up a benchmark.

We also show that HPS v2 is sensitive to algorithmic improvements. Firstly, we evaluate the effectiveness of a straightforward test-time trick of blending the input noise for a diffusion model with a prior image. We also conduct experiments to validate the quantitative improvements of a method proposed by Wu *et al.* [24]. Our experiments show that HPS v2 is sensitive to algorithmic improvements, and can be used to evaluate text-to-image generative models.

Our contributions are as follows: 1) HPD v2, a large-scale, well-annotated dataset for researches of human preferences on images generated by text-to-image generative models. 2) HPS v2, a better human preference prediction model against existing ones, for which we show several example usages

HPS v2: 0.3083 Realistic Vision | HPS v2: 0.2993 Deliberate | HPS v2: 0.2971 MajicMix Realistic | HPS v2: 0.2939 ChilloutMix

Prompt: "A fantasy illustration of a cat standing on a rooftop."

HPS v2: 0.3052 Realistic Vision | HPS v2: 0.3013 Deliberate | HPS v2: 0.2932 DeepFloyd IF | HPS v2: 0.2925 Epic Diffusion

Prompt: "Yoshitaka Amano's painting of a young lion beastman with a white mane, wearing complex fantasy clothing and huge paws, at a medieval market on a windy day."

Figure 2: HPS v2 can be used to automatically evaluate image quality, as well as benchmark text-to-image generative models. The figure illustrates HPS v2 on images generated by 6 open-source text-to-image models.

Table 1: Examples of prompts cleaned by ChatGPT.

Prompts from DiffusionDB [23]	Prompts cleaned by ChatGPT
concept art of translucent glowing curvy buxom brown fairy dancing, renaissance, flowy, melting, round moons, rich clouds, very detailed, volumetric light, mist, fine art, textured oil over canvas, epic fantasy art, very colorful, ornate intricate scales, walls and floor of cave made of skulls, fractal gems, 8 k, hyper realistic	A translucent glowing brown fairy is dancing in a cave made of skulls, surrounded by round moons and rich clouds.
star wars portrait of a robert redford by greg rutkowski, jacen solo, very sad and relucant expression, wearing a biomechanical suit, scifi, digital painting, artstation, concept art, smooth, artstation hq.	A digital painting of Star Wars character Jacen Solo in a biomechanical suit with a sad expression.

to demonstrate its sensitivity and accuracy. 3) A fair, stable and easy-to-use set of evaluation prompts for text-to-image generative models.

2 Related Work

2.1 Text-to-image generative models

Text-to-image generative models enable users to create images based on textual input. DALL·E [16] was the first to achieve high-quality open-domain text-to-image generation. Since then, various architectures and modeling have been explored to enhance image quality. These include autoregressive

models [16, 2], GANs [21], and diffusion models [17, 15, 18]. Among them, diffusion models have demonstrated superior computational efficiency and the ability to produce high-quality samples [15]. However, despite the capability of current text-to-image diffusion models to generate high-fidelity images, they often fail to align well with human preferences. Recent research in this field has attempted to improve the image quality of diffusion models [24, 8], but the evaluation of these improvements remains challenging due to the lack of appropriate metrics and benchmarks. Therefore, there is a need to develop new evaluation metrics and benchmarks to effectively validate algorithmic enhancements.

2.2 Image quality evaluation for generative models

Inception Score (IS) [19] and Fréchet Inception Distance (FID) [5] are widely used in the evaluation of image generative models. However, they do not correlate well with human preferences for images generated by recent text-to-image models [7, 24]. Several recent works [24, 25, 7] propose to finetune visual-language models (VLM) on human choices on images generated with the same prompt, and use finetuned VLMs to serve as proxies for human evaluation, but these models are tuned exclusively on images from Stable Diffusion [17] and its variants, their generalization capabilities are not yet validated.

2.3 Image quality dataset

AVA [11] is a dataset of human ratings given to photographs. However, real images do not contain artifacts that often occur in generated images. Similar to AVA, Simulacra [13] collects human ratings, but instead for images generated from text. We empirically find that rating and preference are two related but different type of data. Jointly training a preference predictor with rating data does not help much with preference prediction. Our HPD v2 includes images from a wide range of sources, enabling a more comprehensive evaluation of these preference prediction models. Besides the diversity of image, there are subtle differences between HPD v2 and these works. HPD v1 [24] is collected without directly prompting the users to choose according to their preference, so the data bias is hard to be measured. ImageReward [25] introduces a dataset collected with a similar methodology like us, but the text prompts are directly taken from DiffusionDB [23], which suffers from severe bias. It is also a smaller dataset compared to ours. Pick-a-Pic [7] is a preference dataset at the similar scale as us. However, the preferences are collected from prompt writers, while ours are collected by professional annotators. It is an interesting topic to study the difference between the two kinds of annotation.

3 HPD v2

We present Human Preference Dataset v2 (HPD v2), a large-scale, cleanly-annotated dataset of human preferences for images generated from text prompts. Our dataset comprises groups of images generated by different models using the same prompt, with human annotators ranking the images in each group based on their preference. Each annotated group of images are generated from the same text prompt and the images are ranked by human annotators, resulting in pairwise binary choices in each group. In total, the dataset contains 798k pairs of binary preference choices for 430k images. We describe the collection pipeline for HPD v2, which includes prompt collection (Sec.3.1), image collection (Sec.3.2), and preference annotation (Sec.3.3). Detailed statistics about the dataset are presented in Sec.3.4.

3.1 Prompt collection

Prompts in HPD v2 are sourced from COCO Captions [1] for realistic content and DiffusionDB [23] for unrealistic content. Prompts from DiffusionDB [23] are firstly cleaned with ChatGPT to reduce bias.

Advantage and issues of prompts in DiffusionDB. DiffusionDB [23] is a valuable source of text prompts for our dataset, because the prompts are written by human users with genuine intention of creating an image. However, there are several issues preventing us from directly using them in our dataset. Firstly, the style words are biased. Certain style words appears frequently, such as artist name "Greg Rutkowski", art platform "ArtStation". The effectiveness of these words is highly dependent

Table 2: Image sources of HPD v2.

Source	Split	# Param.	Type	# images
CogView2	train & test	24B	Autoregressive	73697
DALL·E 2	train & test	3.5B	Diffusion	101869
GLIDE	test	0.94B	Diffusion	400
SD v1.4	train & test	0.89B	Diffusion	101869
SD v2.0	train & test	0.89B	Diffusion	101869
LAFITE	test	0.75B	GAN	400
VQ-GAN+CLIP	test	0.73B	GAN	400
VQ-Diffusion	test	0.37B	Diffusion	400
FuseDream	test	0.35B	GAN	400
COCO Captions	train & test	-	-	28272

on the training data of Stable Diffusion, and may not be suitable for a model trained with different data. This bias can lead to unfair model evaluation, which will be discussed in Sec.5. Another issue is about the organization of prompts in DiffusionDB. They often follow a specific organization of description plus several style words, where the style words can conflict with each other. For example, consider the first prompt in Tab.1. The use of both "melting" and "volumetric light" to describe an artwork could be seen as contradictory, as melting implies a lack of structure while volumetric light suggests a defined form. We find that about 11% of the prompts are conflict by asking ChatGPT. This structural bias may confuse annotators and introduce noise in the annotation.

Prompt cleaning. To address the above issues, we utilize ChatGPT to resolve bias in the prompts. In Tab. 1, we show examples of prompts before and after cleaning. The output prompts are clearly written in one sentence with less style words, which are easier for our annotators to understand. After cleaning, the average prompt length (counted by number of words) decreases from 25.9 to 16.7. The instruction for ChatGPT is elaborated in Appendix. A.

3.2 Image collection

Tab. 2 summarizes image sources of HPD v2, which contains images from 9 recent text-to-image generative models and COCO Captions. Specifically, each group of images for ranking is generated with text-to-image generative models conditioned on the same collected prompt. For prompts from COCO Captions, we add the corresponding ground truth image from COCO to the group as well. To validate the generalization capability of a preference prediction model, images from all 9 models are collected for the test set of HPD v2. However, for the training set, images from only 4 models are collected to test their generalization capability towards other models.

The dataset comprises images from various sources, including models of different architectures, scales, as well as real photos from COCO Captions [1], resulting in a high degree of diversity. This diversity allows for a comprehensive evaluation of a preference prediction model's generalization capability and facilitates the training of a more generalizable model. For information on the inference parameters of each model, please refer to Appendix. B.

3.3 Preference annotation

We hire a team of 57 contractors to label our data, 50 of whom are responsible for actual image ranking, while the remaining 7 are quality control checkers. To ensure consistent annotation, we provide workers with an annotation document for reference, which is elaborated in Appendix. C. Each image group in the training set is randomly assigned to one annotator for ranking, while 10 annotators are assigned to groups in the test set. Annotations are randomly sampled for checkers to review.

3.4 Statistics

HPD v2 comprises a test split and a training split. The test split consists of 400 groups of images, with each group containing 9 images generated from a text-to-image generative model. For "Photo" style, we add one more image from the COCO Captions dataset [1] (thus 10 in total). Each group

in the test set is annotated by 10 distinct annotators to ensure more stable evaluation and to enable the study of diversity in human preference, resulting in 153,000 binary comparisons. The training split of HPD v2 contains 107,515 groups of images, each containing 4 images from 4 models (or real image), ranked by one annotator, which corresponds to 645,090 human binary comparisons. 96.5% of prompts in the training set are unique.

4 Human Preference Score v2

In this section, we introduce the training of Human Preference Score v2 (HPS v2), and present a comprehensive evaluation of existing preference prediction models on the test set of HPD v2 and other related test data [25, 7].

4.1 Training

HPS v2 is trained by finetuning the CLIP model on HPD v2. Each instance in the training set contains a pair of images $\{x_1, x_2\}$ with prompt p, which is labeled with $y = [1, 0]$ if image x_1 is preferred over x_2, otherwise $y = [0, 1]$. The CLIP model can be viewed as a score function s that computes the similarity between prompt p and image x:

$$s_\theta(p, x) = \frac{\mathrm{Enc}_{\mathrm{txt}}(p) \cdot \mathrm{Enc}_{\mathrm{img}}(x)}{\tau}, \tag{1}$$

where τ is the learned temperature scalar of the CLIP model, and θ is the parameters in CLIP. The predicted preference $\hat{y}_i$ is calculated as:

$$\hat{y_i} = \frac{\exp(s_\theta(p, x_i))}{\sum_{j=1}^{2} \exp(s_\theta(p, x_j))}. \tag{2}$$

θ is optimized by minimizing the KL-divergence:

$$L_{\mathrm{pref}} = \sum_{j=1}^{2} y_i(\log y_i - \log \hat{y_j}). \tag{3}$$

We finetune CLIP-H [14] by optimizing L_{pref} on HPD v2 for 4,000 steps with the AdamW optimizer [6, 10], a learning rate of 3.3×10^{-6}, a weight decay of 0.35, a batch size of 128 and a warm-up period of 500 steps, following a cosine learning rate schedule. We train the last 20 layers of the CLIP image encoder, and the last 11 layers of the CLIP text encoder. Hyper-parameters are determined by Bayesian optimization with the target of accuracy on HPD v2 test set.

4.2 Experiments

In Tab. 3, we report the accuracy of the baseline models on test sets of ImageReward [25] and HPD v2, the consistency between human annotators, and the performance of HPS v2. Different from the test set of ImageReward [25] that only contain images generated by Stable Diffusion, the images of HPD v2 test set cover a much wider range of image distribution, as demonstrated in Tab. 2, and is thus more challenging. HPS v2 exhibits a better accuracy on both benchmarks, demonstrating its strong capability of generalization. It should be noted that human preferences are generally diverse. However, the average preference of multiple humans can reflect humans' general tendency. Since HPS v2 is trained on preferences from many annotators, it is able to learn humans' average preference, which can exceed a single person's accuracy.

5 Prompts for evaluating text-to-image generative models

As a model's output is dependent on the input prompt, HPS v2 alone is not sufficient for a comprehensive evaluation. In this section, we first analyze existing prompts for evaluation and then introduce our own choice of prompts.

DrawBench is a list of prompts proposed in Imagen [18] to evaluate text-to-image generative models. It is a collection of 200 prompts organized into 11 categories, designed to test various aspects of

Table 3: Preference prediction accuracy on ImageReward test set and HPD v2 test set.

Model	ImageReward	HPD v2
Aesthetic Score Predictor [22]	57.4	72.6
ImageReward [25]	65.1	70.6
HPS [24]	61.2	73.1
PickScore [7]	62.9	79.8
Single Human	65.3	78.1
HPS v2	**65.7**	**83.3**

Table 4: HPS v2 benchmark. We also provide HPS v2 evaluated on DrawBench for reference.

	HPS v2				
Model	Animation	Concept-art	Painting	Photo	DrawBench [18]
GLIDE [12]	$0.2334 \pm 1.98 \times 10^{-3}$	$0.2308 \pm 1.74 \times 10^{-3}$	$0.2327 \pm 1.78 \times 10^{-3}$	$0.2450 \pm 2.90 \times 10^{-3}$	0.2505
LAFITE [27]	$0.2463 \pm 1.01 \times 10^{-3}$	$0.2438 \pm 0.87 \times 10^{-3}$	$0.2443 \pm 1.55 \times 10^{-3}$	$0.2581 \pm 2.13 \times 10^{-3}$	0.2523
VQ-Diffusion [4]	$0.2497 \pm 1.86 \times 10^{-3}$	$0.2470 \pm 1.49 \times 10^{-3}$	$0.2501 \pm 1.45 \times 10^{-3}$	$0.2571 \pm 2.22 \times 10^{-3}$	0.2544
FuseDream [9]	$0.2526 \pm 1.25 \times 10^{-3}$	$0.2515 \pm 1.07 \times 10^{-3}$	$0.2513 \pm 1.83 \times 10^{-3}$	$0.2557 \pm 2.48 \times 10^{-3}$	0.2572
Latent Diffusion [17]	$0.2573 \pm 1.25 \times 10^{-3}$	$0.2515 \pm 1.40 \times 10^{-3}$	$0.2525 \pm 1.78 \times 10^{-3}$	$0.2697 \pm 1.83 \times 10^{-3}$	0.2617
CogView2 [2]	$0.2650 \pm 1.29 \times 10^{-3}$	$0.2659 \pm 1.19 \times 10^{-3}$	$0.2633 \pm 1.00 \times 10^{-3}$	$0.2644 \pm 2.71 \times 10^{-3}$	0.2617
DALL·E mini	$0.2610 \pm 1.32 \times 10^{-3}$	$0.2556 \pm 1.37 \times 10^{-3}$	$0.2556 \pm 1.12 \times 10^{-3}$	$0.2612 \pm 2.33 \times 10^{-3}$	0.2634
Versatile Diffusion [26]	$0.2659 \pm 1.78 \times 10^{-3}$	$0.2628 \pm 1.45 \times 10^{-3}$	$0.2643 \pm 1.02 \times 10^{-3}$	$0.2705 \pm 2.29 \times 10^{-3}$	0.2677
VQGAN + CLIP [3]	$0.2644 \pm 1.52 \times 10^{-3}$	$0.2653 \pm 0.75 \times 10^{-3}$	$0.2647 \pm 1.11 \times 10^{-3}$	$0.2612 \pm 2.10 \times 10^{-3}$	0.2638
DALL·E 2 [15]	$0.2734 \pm 1.75 \times 10^{-3}$	$0.2654 \pm 1.27 \times 10^{-3}$	$0.2668 \pm 1.56 \times 10^{-3}$	$0.2724 \pm 1.98 \times 10^{-3}$	0.2716
Stable Diffusion v1.4 [17]	$0.2726 \pm 1.56 \times 10^{-3}$	$0.2661 \pm 0.82 \times 10^{-3}$	$0.2666 \pm 1.43 \times 10^{-3}$	$0.2727 \pm 2.26 \times 10^{-3}$	0.2723
Stable Diffusion v2.0 [17]	$0.2748 \pm 1.74 \times 10^{-3}$	$0.2689 \pm 0.76 \times 10^{-3}$	$0.2686 \pm 1.20 \times 10^{-3}$	$0.2746 \pm 1.98 \times 10^{-3}$	0.2731
Epic Diffusion	$0.2757 \pm 1.63 \times 10^{-3}$	$0.2696 \pm 1.13 \times 10^{-3}$	$0.2703 \pm 0.88 \times 10^{-3}$	$0.2749 \pm 1.92 \times 10^{-3}$	0.2733
Openjourney	$0.2785 \pm 1.45 \times 10^{-3}$	$0.2718 \pm 0.90 \times 10^{-3}$	$0.2725 \pm 1.24 \times 10^{-3}$	$0.2753 \pm 1.78 \times 10^{-3}$	0.2744
MajicMix Realistic	$0.2788 \pm 1.97 \times 10^{-3}$	$0.2719 \pm 0.94 \times 10^{-3}$	$0.2722 \pm 1.49 \times 10^{-3}$	$0.2764 \pm 1.76 \times 10^{-3}$	0.2747
ChilloutMix	$0.2792 \pm 1.31 \times 10^{-3}$	$0.2729 \pm 0.90 \times 10^{-3}$	$0.2732 \pm 1.54 \times 10^{-3}$	$0.2761 \pm 1.95 \times 10^{-3}$	0.2747
DeepFloyd-XL	$0.2764 \pm 1.08 \times 10^{-3}$	$0.2683 \pm 1.37 \times 10^{-3}$	$0.2686 \pm 1.31 \times 10^{-3}$	$0.2775 \pm 1.71 \times 10^{-3}$	0.2764
Deliberate	$0.2813 \pm 1.35 \times 10^{-3}$	$0.2746 \pm 0.98 \times 10^{-3}$	$0.2745 \pm 1.11 \times 10^{-3}$	$0.2762 \pm 2.01 \times 10^{-3}$	0.2773
Realistic Vision	$0.2822 \pm 1.33 \times 10^{-3}$	$0.2753 \pm 0.93 \times 10^{-3}$	$0.2756 \pm 1.24 \times 10^{-3}$	$0.2775 \pm 2.26 \times 10^{-3}$	0.2777
Dreamlike Photoreal 2.0	$0.2824 \pm 1.43 \times 10^{-3}$	$0.2760 \pm 0.85 \times 10^{-3}$	$0.2759 \pm 1.10 \times 10^{-3}$	$0.2799 \pm 2.45 \times 10^{-3}$	0.2788

a model's capabilities. However, there is currently no consensus on how to evaluate text-to-image generation models using DrawBench, aside from user studies, which can be difficult to carry out evaluation consistently across different models. With the help of HPS v2, DrawBench comparisons can now be conducted consistently without relying on user studies. The results are presented in Table 4.

COCO Captions dataset. Captions from COCO Captions are used for evaluation in Pick-a-Pic [7], and not suggested because they are descriptions for photographs, not for "fiction". While it is important to consider prompts of "fiction", we include prompts from COCO Captions in our benchmark for "Photo" style, to ensure a comprehensive evaluation.

User-written prompts. DiffusionDB [23] is a database that contains a wide range of user-written prompts and generated images. However, as discussed in Sec. 3.1, a significant portion of the prompts in the database are biased towards certain styles. For instance, around 10.4% of the prompts in DiffusionDB include the name "Greg Rutkowski", 21.9% include "artstation", highlighting the importance of evaluating models on diverse and unbiased prompts to obtain a more accurate assessment of their capabilities. By cleaning via ChatGPT, the portion of prompts that contain "Greg Rutkowski" and "artstation" decreases to 2.7% and 3.1%, which is less biased than prompts in DiffusionDB.

Our evaluation prompts. Based on the observations above, we recommend evaluating models separately for different styles using less biased prompts obtained from ChatGPT, as described in Sec. 3.1. To this end, we propose a set of evaluation prompts that involves testing a model on a total of 3200 prompts, with 800 prompts for each of the following styles: "Animation", "Concept-art", "Painting", and "Photo". The prompts for the first three styles are collected using the method outlined in Sec.3.1, while the "Photo" style uses prompts from COCO Captions [1]. For each style, 800 prompts are divided into groups of 80, and HPS v2 is computed on each group. The mean and standard deviation of the HPS v2 scores are then reported. We chose the amount of 800 to ensure that HPS v2 is statistically stable across all evaluated models, while also avoiding excessive computational overhead.

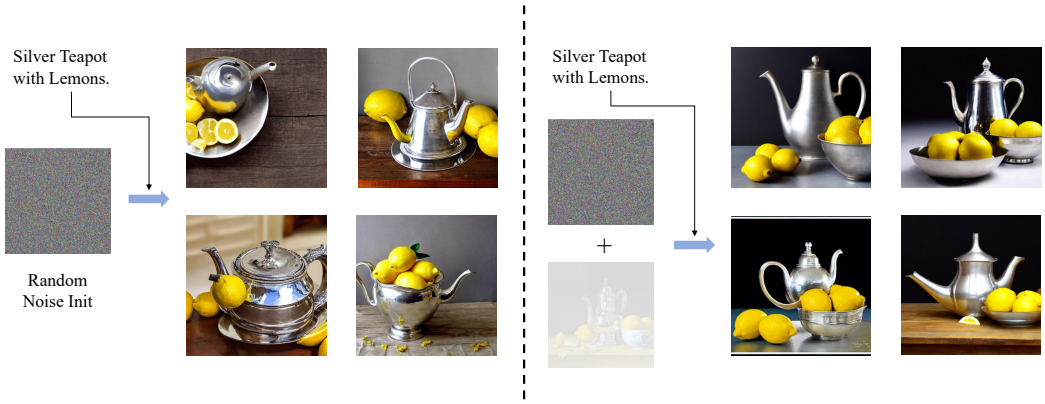


Figure 3: Left: the standard denoising inference. Right: blending the random initial noise with a prior image. The input random noise has strong influence on the layout of the generated image.

Benchmark. In Tab. 4, we report our benchmark consisting a relatively complete list of recent from the academia, community and industry. We are able to observe an obvious trend that models that are popular in the community are consistently out-performing models from the academic. Inference details will be elaborated in Appendix. B.

6 Example usages of HPS v2

We show 2 example usages of HPS v2 in this section to show its sensitivity and accuracy.

6.1 Retrieval Initialization

Method. Many recent works [20] suggest that the latent feature initialization has a great influence on the output image quality of Stable Diffusion. We observe that this may relate to a misalignment between the noise schedule of training and inference process. We find that when training Stable Diffusion, there isn't a timestep at which the input latent feature is fully random, but at inference time the image is recovered from random noise, which is inconsistent with the training stage. This implies that the inference of Stable Diffusion can be potentially improved by initializing with better latent features. With this observation, we try a straightforward idea of blending the latent feature of a reference image with random noise to initialize the latent features.

Given a prompt, a prior image is generated by a powerful community model Dreamlike Photoreal 2.0, and then the image is encoded into latent space with noise level similar to the training stage. The noisy latent feature is used to initialize the inference of Stable Diffusion, which is previously sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ without prior.

Results. We verify the effectiveness of this method by comparing it with the default initialization of Stable Diffusion. Tab.5 demonstrates that the improvement brought by initializing via retrieved images can be verified by HPS v2. Visualizations in Fig.3 also show that the noise initialization has a strong impact on the global layout of the generated image, thus randomly initializing it may lead to undesired image layouts.

6.2 Evaluating adapted model in HPS v1

Wu et al.[24] improves Stable Diffusion to better align it with human preference, as introduced in HPS v1. However, the improvement is only validated through user studies. There are two reasons that prevent the authors from evaluating via HPS v1. Firstly, the bias of HPS v1 and the evaluation prompts are under-studied, so HPS v1 is not ready for serving as an evaluation metric. Secondly, HPS v1 also incorporates the training of the adapted model. With HPS v2, we are now able to quantify the actual improvement of the method, as shown in Tab.5.

Table 5: Example usages of HPS v2.

Method	Animation	Concept-art	Painting	Photo
Stable Diffusion v1.4 [17]	0.2726	0.2661	0.2666	0.2727
Retrieval initialization	0.2739	0.2659	0.2671	0.2746
Human-aligned tuning [24]	0.2780	0.2716	0.2724	0.2760

7 Limitations

There are several potential limitations about this work. The prompts are sourced from DiffusionDB [23] and COCO Captions [1]. Although DiffusionDB provides a massive collection of prompts that reflects community users' demands, there are indeed some overlooked topics, such as logo and graphic design. These demands also play a crucial role in design industries and have unique criterion or preference compared to general image creation. Also, this dataset is annotated by 57 annotators, and thus it may suffer from bias of them.

8 Conclusion

In this work, we design a less biased pipeline for collecting prompts and images for human preference annotations. With collected data, we present Human Preference Dataset v2, a human preference dataset containing 798k carefully annotated comparison pairs, enabling the training of Human Preference Score v2, the state-of-art preference prediction model, to better align text-to-image generation models' evaluation with human values and judgments. We also provide a more stable and fair group of prompts for text-to-image generative models' evaluation, and set up a benchmark that compares a wide range of recent models. Finally, we illustrate the accuracy of HPS v2 by demonstrating some example usages. There are still many opportunities for further advancement and expansion based on our HPD v2 and HPS v2. We hope our work can inspire and facilitate future researches in this area.

References

[1] 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.

[2] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

[3] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

[4] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 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, 2022.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] 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.

[8] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[9] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[11] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012.

[12] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021.

[13] John David Pressman, Katherine Crowson, and Simulacra Captions Contributors. Simulacra aesthetic captions. Technical Report Version 1.0, Stability AI, 2022. url https://github.com/JD-P/simulacra-aesthetic-captions .
[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.
[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.
[16] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021.
[17] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.
[18] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022.
[19] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
[20] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. It is all about where you start: Text-to-image generation with seed selection. *arXiv preprint arXiv:2304.14530*, 2023.
[21] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023.
[22] 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. *arXiv preprint arXiv:2210.08402*, 2022.
[23] 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.
[24] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference, 2023.
[25] 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, 2023.
[26] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022.
[27] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Sec. 4.1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In Tab. 4, we report the mean and standard deviation of 10 runs to demonstrate the stability and reliability of our benchmark.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The resource consumption will be reported in our GitHub repository.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We will release our dataset and pre-train model.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Please see Sec. 3.3.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We will show our instructions given to the workers in the supplemental material.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

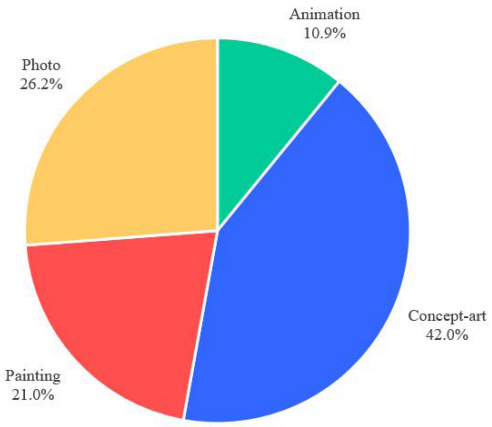


Figure 4: Proportions of category “Animation”, “Concept-art”, “Painting” and “Photo”

A Prompt Cleaning

In this section we show details in prompt cleaning, categorization and annotation. As illustrated in Sec. 3.1, we task ChatGPT to clean and categorize prompts by posing the following questions.

> -I will give you a description about an image.Remove modifiers from text that have nothing to do with the main content of the image, for example resolution, sharpness, light, image quality, authors and online platform, and describe it succinctly in one sentence.
>
> -Next, I will use text to describe a picture. Please reply to me according to the following rules:
> 1. If the picture belongs to the style of ’paintings’ ,reply only with ’paintings’ ;
> 2. If the picture belongs to the style of ’anime and cartoon’ ,reply only with ’anime and cartoon’;
> 3. If the picture belongs to the style of ’real photo’ ,reply only with ’real photo’ ;
> 4. If the picture belongs to the style of ’concept-art’ ,reply only with ’concept-art’ ;
> 5. If the picture doesn’t belong to any styles of above ,reply only with ’others’;
> You must reply with only on word.

Even though prompts of “Photo” category in HPD v2 are from COCO Captions [1], we retain “Photo” in the classification process to mitigate the potential mistakes made by ChatGPT. The category distribution of HPD v2 is illustrated in Fig. 4.

Additionally, when processing data from DiffusionDB [23], we apply a filter to remove prompts with a NSFW threshold of 0.4. During the subsequent image generation step, we refine the selection of prompts by subjecting them to a stricter safety checker of DALL·E 2 [15]. This helps to further eliminate inappropriate vocabularies.

B Generation Details

We provide image generation details for each model when constructing the HPS v2 benchmark. The images are also released with HPD v2.

Stable Diffusion. For Stable Diffusion v1.4 and v2.0, we generate images of size 512×512 using a guidance scale of 7.5 for 50 steps with DDIM sampler.

GLIDE. Images of size 256×256 are generated from the official mini-glide model with a guidance scale of 3.0 for 27 steps with DDIM sampler.

(a) (b)

Figure 5: Prompt: A pair of skis standing up against a gate.

LAFITE. Images of size 256×256 are generated following default configuration using model[*].

VQ-Diffusion. Images of size 256×256 are generated with a guidance scale of 5.0 and 100 sampling steps with VQ-Diffusion sampler using model[†].

FuseDream. Images of size 512×512 are generated with 1000 initialization iterations, 1000 optimization iterations and basis image number is set to 5.

Latent Diffusion. Images are generated from the official LDM model with the DPM Solver sampler, 25 sampling steps, $\eta = 0.3$, and a guidance scale of 6.0.

CogView2. Images of size 512×512 are generated with default hyper-parameters. For category "animation", "concept-art", "painting" and "photo" in HPD v2, we set the "style" in Cogview2 to "none", "none", "oil" and "photo" respectively.

DALL·E 2. Images are directly collected by requesting the official API[‡].

Versatile Diffusion. Images are generated from the official model using the DPM Solver sampler, 25 sampling steps, and a guidance scale of 7.5.

VQGAN+CLIP. Images of size 512×512 are generated with 500 iterations and Adam optimizer with a learning rate of 0.1.

Community Models based on Stable Diffusion (Epic Diffusion, Openjourney, MajicMix Realistic, ChilloutMix, Deliberate, Realistic Vision, Dreamlike Photoreal 2.0) Images of size 512×512 are generated from the publicly released model with the DPM Solver sampler, 25 sampling steps, and a guidance scale of 7.5.

DeepFloyd-XL Stages 1 and 2 are set to use DDPM noise schedulers with 25 sampling steps and the default samplers and guidance scales. Then the generated 256×256 images are resized to 192×192. For the upsampling stage, we use Stable Diffusion upscaler, with the DPM Solver sampler and 25 steps to generate 768×768 images.

DALL·E mini We use the fast and minimal reproduction of the model[§] and all hyper-parameters are set to default except $grid\_size = 1$.

C Annotation

To ensure the quality of annotation, we provide a basic guidance document for our annotators:
Basic rules and regulations

[*]https://drive.google.com/file/d/1tMD6MWydRDMaaM7iTOKsUK-Wv2YNDRRt/view?usp=sharing
[†]https://huggingface.co/microsoft/vq-diffusion-ithq
[‡]https://openai.com/
[§]https://huggingface.co/kuprel/min-dalle

We will provide you groups of images for ranking, which consists of images generated from different AI models. Please consider the prompts and rank images from the perspectives of universal and personal aesthetic appeal. This task mainly involves two aspects: text-image alignment and image quality. Although we encourage and value personal preference, it's important to consider the following fundamental principles when balancing the two aspects or facing a dilemma:

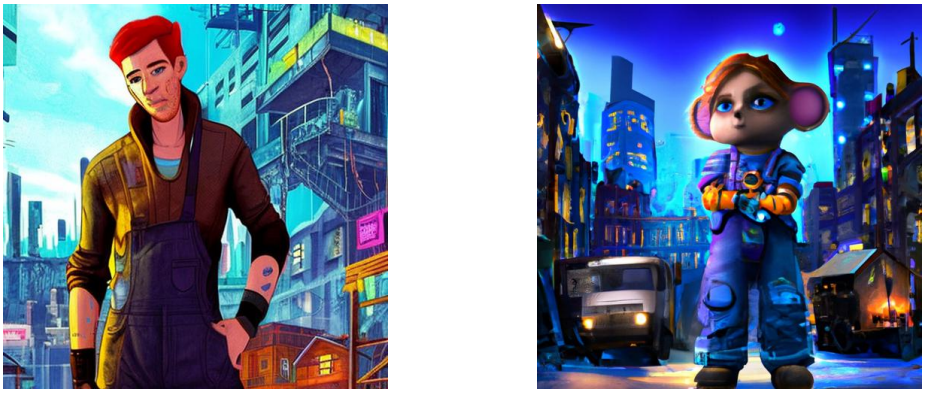
1. When Image (A) surpasses Image (B) in terms of aesthetic appeal and fidelity, or Image (B) suffers from severe distortion and blurriness, even if Image (B) aligns better with the prompt, Image (A) should take precedence over Image (B). For example, in Fig. 5, Fig. 5(b) lacks clear outlines and details of the ski board, resulting in an unattractive appearance and significant blurriness. However, Fig. 5(b) exhibits excellent fidelity and quality. Therefore, the ordering of the figures should place Fig. 5(a) before Fig. 5(b).

(a) (b)

Figure 6: Prompt: A cat with two horns on its head.

Similarly, in Fig. 6, although in Fig. 6(a) the required horns are mistakenly generated as elf ears, Fig. 6(b) suffers severe structural issues, repetitive generation problems, and blurriness. Therefore, the ordering of the figures should place Fig. 6(a) before Fig. 6(b).

2. When facing a dilemma that images are relatively similar in terms of aesthetics and personal preference, please carefully read and consider the prompt for sorting based more on the text-image alignment. For example, if you cannot make a choice based on personal preference, as in Fig. 7, please pay attention to the description, which refers to a mouse mechanic. Therefore, the ordering of the figures should place Fig. 7(b) before Fig. 7(a).



(a) (b)

Figure 7: Prompt: A ginger haired mouse mechanic in blue overalls in a cyberpunk scene with neon slums in the background.

Table 6: Pairwise Preference Prediction Accuracy of HPSv2

Model	SD v1.4	SD v2.0	VQ-Diffusion	LAFITE	GLIDE	CogView2	VQ-GAN+CLIP	DALL·E2	COCO
FuseDream	87.3%	92.5%	67.8%	79.3%	90.2%	85.5%	51.5%	90.3%	92.0%
SD v1.4	-	72.0%	90.3%	92.0%	96.5%	67.5%	70.5%	78.5%	85.0%
SD v2.0	-	-	93.3%	97.0%	98.8%	76.5%	83.0%	73.8%	79.0%
VQ-Diffusion	-	-	-	80.8%	88.0%	90.0%	51.5%	91.3%	93.0%
LAFITE	-	-	-	-	82.8%	92.0%	73.8%	92.8%	95.0%
GLIDE	-	-	-	-	-	96.8%	84.8%	96.8%	96.0%
CogView2	-	-	-	-	-	-	73.0%	79.0%	84.0%
VQ-GAN+CLIP	-	-	-	-	-	-	-	79.5%	92.0%
DALL·E 2	-	-	-	-	-	-	-	-	74.0%

3. It is crucial to pay special attention to the capitalized names, as these names may lead to misunderstandings during the machine translation process. If there is any incorrectly translated proprietary term or content you are not familiar with, we recommend you to search for sample images and explanations online.

D Pairwise Accuracy

To further illustrate the accuracy of HPS v2, we evaluate its agreement with human choices between all pairs of generative models in test split. As shown in Tab. 6, the predictions of our model align closely with human choices.

E Datasheet

E.1 Motivation

Why was the dataset created?

The dataset was created to facilitate future academic Computer Vision research about human aesthetic preference.

Who created this dataset (*e.g.*which team, research group) and on behalf of which entity (*e.g.*company, institution, organization)?

The dataset was created by researchers at MMLab, The Chinese University of Hong Kong and Centre for Perceptual and Interactive Intelligence (CPII).

E.2 Composition

What do the instances that comprise the dataset represent (*e.g.*documents, photos, people, countries)? Are there multiple types of instances? (*e.g.*movies, users, ratings; people, interactions between them; nodes, edges)

The instances are prompts and generated images, along with human preference choices among the images generated by the same prompt.

Are relationships between instances made explicit in the data (*e.g.*social network links, user/movie ratings, etc.)?

Only in benchmark. In training and test split, due to personal privacy concerns, we didn't include any annotator information in our dataset. Therefore, it's not possible to query data from the same annotators. In another aspect, we didn't provide the model from which the image was generated, since during training we are more interested in human preference rather than sources of images. However, in benchmark, images from the same model can be directly retrieved using their respective IDs.

How many instances are there? (of each type, if appropriate)?

There are 789,090 instances in the dataset, standing for 789,090 comparisons between image pairs. Specifically, our dataset has 433,760 images and 107,515 prompts. The benchmark contains 3200 unique prompts without paired images.

What data does each instance consist of? "Raw" data (*e.g.* unprocessed text or images) or Features/attributes? Is there a label/target associated with instances? If the instances related to people, are sub-populations identified (*e.g.* by age, gender, etc.) and what is their distribution?

Each instance consists of two images, one prompt and one human choice.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (*e.g.* because it was unavailable). This does not include intentionally removed information, but might include, *e.g.* redacted text.

Yes, we omit the specific parameters for generating the images, such as diffusion steps and guidance scale. They are omitted because we are more interested in the users' preference about the generated images, rather than how they are created. Instead, we illustrate the hyper-parameters used for generation in appendix, see Sec. B.

Is everything included or does the data rely on external resources?

The dataset is self-contained.

Are there recommended data splits and evaluation measures? (*e.g.* training, development, testing; accuracy or AUC)

In our experiments, we use a training set of 645,090 instances and a validation set of 153,000 images, which will be made public. We recommend using accuracy (%) with one decimal place.

Are there any errors, sources of noise, or redundancies in the dataset?

There is potential noise in the dataset. Since human preference is highly diverse, in most cases, it's hard to tell whether the divergence in human choice is due to noise or disagreement of aesthetic criterion.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (*e.g.* websites, tweets, other datasets)?

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (*e.g.* data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)?

No, both prompt sources (DiffusionDB and COCO Captions) and image generative models are publicly available.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

We retained the prompts with NSFW scores below a threshold and accepted by DALL·E2. However, there may still be an extremely small portion of inappropriate content that were not removed by our policy.

Does the dataset relate to people?

Yes, the choices are made by human annotators.

Does the dataset identify any subpopulations (*e.g.*by age, gender)?

No.

Is it possible to identify individuals (*i.e.*one or more natural persons), either directly or indirectly (*i.e.*in combination with other data) from the dataset?

No.

Does the dataset contain data that might be considered sensitive in any way (*e.g.*data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

The dataset may contain sensitive data, because the prompts are sourced from DiffusionDB, which are collected from users. Thus, it may contain sensitive information, such as public figures and religious beliefs.

What experiments were initially run on this dataset? Have a summary of those results.

We firstly collected the test split to evaluate existing preference prediction models and found that they do not align well with human choices. Details are illustrated in Tab. 3. Based on this, we collected the large scale training split of HPD v2 and train a more aligned preference prediction model.

E.3 Data Collection Process

How was the data associated with each instance acquired?

Each instance contain two images generated from the same prompt, and a human choice annotation.

What mechanisms or procedures were used to collect the data (*e.g.*hardware apparatus or sensor, manual human curating, software program, software API)?

We use ChatGPT API provided by OpenAI to process prompts from DiffusionDB and an annotation system.

If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*deterministic, probabilistic with specific sampling probabilities)?

The dataset is not a sample of a larger set.

Who was involved in the data collection process (*e.g.*students, crowd-workers, contractors) and how were they compensated (*e.g.*how much were crowdworkers paid)?

We hire 57 annotators for human preference annotation, but they prefer not to disclose their salary information due to personal privacy reasons.

Over what time-frame was the data collected?

The dataset was collected between 2023-03-15 and 2023-06-03.

Were any ethical review processes conducted (*e.g.*by an institutional review board)?

No official process is conducted.

Does the dataset relate to people?

No. Although crowdsourcing was used for annotation, we didn't perform any kind of intervention or interaction during data collection and didn't collect any personal information or biospecimens of

annotators. It's not possible to identify individuals from our dataset. In addition, according to the definition of "Human Subjects Research" given by NIH(National Institutes of Health of USA)[¶], our research is not regarded as "research with human subjects".

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (*e.g.*websites)?

We collected human choices directly from annotators in question.

Has an analysis of the potential impact of the dataset and its use on data subjects (*e.g.*a data protection impact analysis)been conducted?

No analysis has been conducted.

E.4 Data Preprocessing

What preprocessing/cleaning was done? (*e.g.*discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

We conducted prompt cleaning process using ChatGPT to remove biased functional words. Please see Sec. A and Sec. 3.1 for details.

E.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

As described in the paper, this dataset has been used for training HPS v2.

Is there a repository that links to any or all papers or systems that use the dataset?

No.

What (other) tasks could the dataset be used for?

It can be used for tasks related to human preference on generated images.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Yes. As discussed in Sec. 7, there might be bias induced by the limited number of annotators.

Are there tasks for which the dataset should not be used?

No.

E.6 Data Distribution

Will the dataset be distributed to third parties outside of the entity (*e.g.*company, institution, organization) on behalf of which the dataset was created? If so, please provide a description

Yes. Anyone can access the dataset via internet.

How will the dataset be distributed? (*e.g.*tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

We will provide a download link in a GitHub repository.

[¶]https://grants.nih.gov/policy/humansubjects/research.htm

When will the dataset be distributed?

When the paper is accepted.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

We will provide a terms of use agreement with the dataset. The dataset as a whole will be distributed under a non-commercial license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown.

E.7 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The authors of this paper are maintainers of this dataset.

How can the owner/curator/manager of the dataset be contacted (*e.g.*email address)?

By email: wuxiaoshi@link.cuhk.edu.hk.

Is there an erratum?

At this time, we are not aware of errors in our dataset. However, we will create an erratum as errors are identified.

Will the dataset be updated? If so, how often and by whom? How will updates be communicated? (*e.g.*mailing list, GitHub)

The dataset will be updated by the authors on an at-will basis (but no more than once a month).

Will older versions of the dataset continue to be supported/hosted/maintained?

N/A

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

There will not be a mechanism to build on top of the dataset.