Human Preference Score: Better Aligning Text-to-Image Models with Human Preference

Xiaoshi Wu[1], Keqiang Sun[1], Feng Zhu[2], Rui Zhao[2,3], Hongsheng Li[1,4,5]
[1]Multimedia Laboratory, The Chinese University of Hong Kong
[2]SenseTime Research [3]Qing Yuan Research Institute, Shanghai Jiao Tong University
[4]Centre for Perceptual and Interactive Intelligence (CPII) [5]Shanghai AI Laboratory
{wuxiaoshi@link, kqsun@link, hsli@ee}.cuhk.edu.hk, {zhufeng, zhaorui}@sensetime.com

Abstract

Recent years have witnessed a rapid growth of deep generative models, with text-to-image models gaining significant attention from the public. However, existing models often generate images that do not align well with human preferences, such as awkward combinations of limbs and facial expressions. To address this issue, we collect a dataset of human choices on generated images from the Stable Foundation Discord channel. Our experiments demonstrate that current evaluation metrics for generative models do not correlate well with human choices. Thus, we train a human preference classifier with the collected dataset and derive a Human Preference Score (HPS) based on the classifier. Using HPS, we propose a simple yet effective method to adapt Stable Diffusion to better align with human preferences. Our experiments show that HPS outperforms CLIP in predicting human choices and has good generalization capability toward images generated from other models. By tuning Stable Diffusion with the guidance of HPS, the adapted model is able to generate images that are more preferred by human users. The project page is available here: https://tgxs002.github.io/align_sd_web/.

1. Introduction

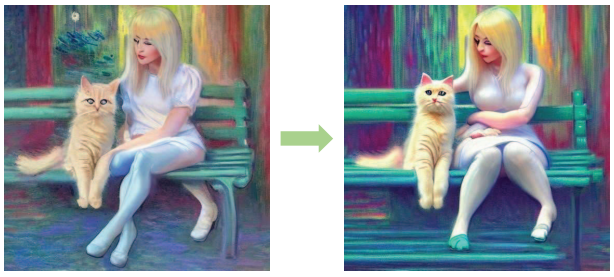
The recent progress in diffusion models [26, 30, 35, 32] has enabled impressive advancements in text-to-image generation, with many models now being deployed in real-world applications such as DALL·E [30] and Stable Diffusion [32]. However, public attention has also highlighted new issues, such as the awkward combinations of limbs and facial expressions of generated persons as shown in Fig. 1. The users usually need to cherry-pick results to avoid these artifacts. In other words, the generated images are misaligned with human preferences.

To further improve the quality of generated images, it is essential to track the ability of a model to generate human preferable images. However, it is uncertain whether the existing evaluation metrics, such as Inception Score (IS) [36] and Fréchet inception distance (FID) [13], are correlated with human choices. These metrics perceive an

beautiful blonde woman in the image of a fairy - tale princess in the garden with a wreath in her hands ...

portrait of an old man with greying hair and a wrinkled face, taking a grandfather clock apart, dreary cityscape background ...

Figure 1. Generated images often do not align well with human preferences and intentions. Input prompts are shown below images.



a blonde woman with a ragdoll cat sitting next to each other on a bench, cyberpunk art by monet, trending on cgsociety, retrofuturism, reimagined by industrial light and magic, darksynth, sci - fi

Figure 2. We show that Stable Diffusion v1.4 can be adapted to better align with human preferences and intentions when guided by the proposed human preference classifier. The input prompt is shown below images.

image through a classification-based CNN trained on ImageNet [34], which has been shown to be biased towards image texture rather than general image contents [10], and thus may not align well with human perception. Also, both IS and FID are single-modal evaluation metrics, which do not take user intention into account. Some recent studies [26, 30, 3] use the CLIP [29] model as a proxy for human judgment to evaluate the alignment between generated images and text prompts. The CLIP [29] model is trained on a rich dataset and is believed to capture subtle aspects of human intention better. However, it is uncertain whether CLIP [29] can measure the quality of generated synthetic images, which may not adhere to the same constraints as real images, such as the example shown in Fig. 1.

In this study, we investigate the problem of human preference using a novel, large-scale dataset of human choices on images generated by Stable Diffusion [32] using the same prompt. The dataset comprises 98,807 diverse images generated from user-provided prompts, along with 25,205 human choices. By evaluating on this dataset, we find that the Inception Score (IS) [36], the Fréchet Inception distance (FID) [13] and the CLIP score does not fully match the human choice, which means that the human preference is a missing dimension of image quality that is not well tracked by existing mainstream metrics.

We further train a human preference classifier on this dataset by fine-tuning the CLIP [29] model and define human preference score (HPS) based on it. We validate HPS's alignment with human choices and its generalization capability towards other generative models through user studies. HPS can be utilized to guide generative models toward producing human-preferred images. To this end, we devise a simple yet effective method to adapt Stable Diffusion [32] by LoRA [17] with awareness of human preference. We conduct user studies to validate the effectiveness of our approach. The results show that the adapted model can better capture human intentions, and generate more preferable images, which significantly mitigates the kind of artifact shown in Fig. 1.

Our contributions are as follows: (1) We create a large-scale dataset for studying human preferences. To our best knowledge, this dataset is the first of its kind that contains massive human choices on images generated with the same prompt. (2) We find that human choices cannot be accurately predicted by the existing mainstream evaluation metrics, while it can be better predicted via fine-tuning CLIP on the proposed dataset. (3) We propose a simple yet effective method to guide the Stable Diffusion model toward generating images with better aesthetic quality and better alignment with human intention.

2. Related Works

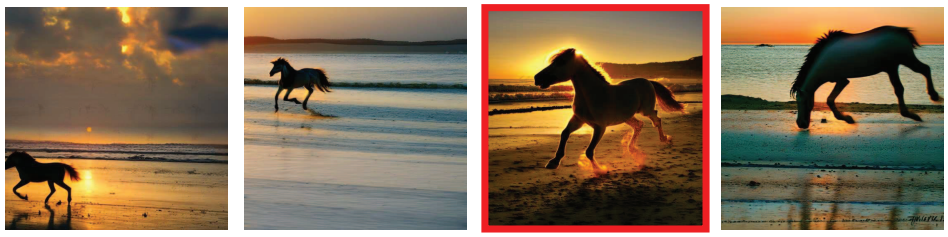
Text-to-image generative models. Text-to-image generative models have long been an active research area. Mansimov *et al*. [23] show that Deep Recurrent Attention Writer (DRAW) [11] can be conditioned on captions to generate novel scene compositions. Generative Adversarial Networks (GANs) improve image fidelity by training a discriminator to provide supervision for the generative model. DALL·E [31] firstly achieves open-domain text-to-image synthesis with the help of massive image-text pairs.

Diffusion models formulate the generative process as the inverse of the diffusion process [41], which was improved by Song and Ermon [42] and Ho *et al*. [15]. Dhariwal *et al*. firstly show the superiority of diffusion models over GANs on image generation. Several following works, including DALL·E 2 [30], GLIDE [26], Imagen [35], ERNIE-ViLG [9, 47], Stable Diffusion [32], bring the magic of text-to-image generation to the public attention. Among these models, Stable Diffusion is an open-source model with an active user community.

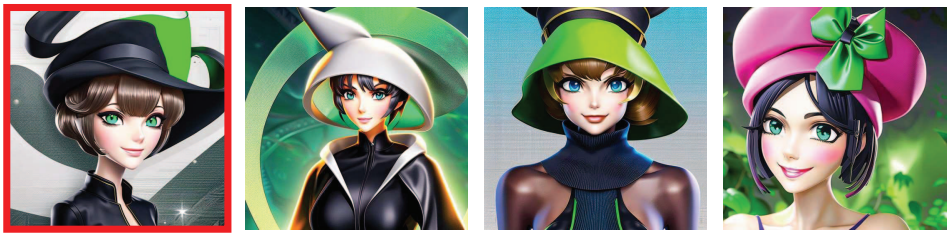
Several recent works improve Stable Diffusion on different aspects. DreamBooth [33] and ELITE [46] explore customizing Stable Diffusion to a certain object. Feng *et al*. [8] propose a training-free method to guide diffusion models for better compositional capabilities. It has been discovered that prompt engineering plays an important role in generating high-quality images. Hao *et al*. [12] devise an automatic prompt engineering scheme via reinforcement learning. Our method focuses on the misalignment between the generated image and human preference, which is orthogonal to the above-mentioned topics.

Datasets of generated images. Datasets of generated images play a vital role in computer vision tasks that has difficulty in ground-truth acquisition, such as optical flow estimation [24, 7, 4, 44, 18, 39, 40]. Thanks to active user communities of text-to-image models, several databases of images generated by diffusion models have been introduced. Lexica (lexica.art) is a large database of images generated by Stable Diffusion and Lexica Aperture. It also provides related information about the image, such as the prompt and guidance scale. However, the database is closed-source and only allows online browsing. DiffusionDB [45] is a large-scale open-source database collected from the Stable Foundation Discord channel, containing the text prompt and parameters for each image. SAC [28] is a dataset of images generated from Stable Diffusion and GLIDE [26], along with user ratings from an aesthetic survey. However, SAC only contains limited user choices compared to our dataset.

Learning from human feedback. Human feedback has long been used in a wide range of deep learning tasks. Christiano *et al*. [6] and Arakawa *et al*. [1] incorporate human feedback into RL training, which is proven to accelerate the model convergence. Krishna *et al*. [20] propose



A horse running on a beach at sunrise, volumetric lighting



The personification of the halloween holiday in the form of a cute girl with short hair ...

Figure 3. Examples of the collected data. The images are generated by Stable Diffusion, with corresponding prompts shown below each row of images. The preferred images are highlighted with red borders. More examples can be found in the appendix.

"socially situated AI", which significantly improves image recognition performance via interacting with human users on Instagram. InstructGPT [27] fine-tunes GPT via a reward function trained on human feedback, establishing the foundation for the success of ChatGPT. [12] and [21] use similar methodology to improve text-to-image models, which are highly related to our work. In [12], this is achieved by augmenting the text prompt. [21] is a concurrent work that focuses more on the exact alignment between text and image, while our work shows that the potential of human feedback is far beyond the exact alignment when the feedback takes into account the aesthetic preference of humans.

3. Human Preference Dataset

In order to get a better understanding of the human preferences on the images generated from prompts, and to improve text-to-image generation quality, we start by collecting a dataset of human choices.

Data collection. We utilize the "dreambot" channel on the Stable Foundation Discord server to gather human choice data. The chat history of these channels is obtained using the DiscordChatExporter [16] tool, which downloads the full chat history of a Discord channel and stores it in JSON format. Among the chat messages, a discernible pattern of interaction is observed, as depicted in Fig. 4, which reveals human preferences. In this pattern, a user initiates a session by sending a text prompt to the bot, which generates several images in response. Then, the user selects a preferred image and sends it back to the bot, along with the original text prompt. The bot will return several refined images. This

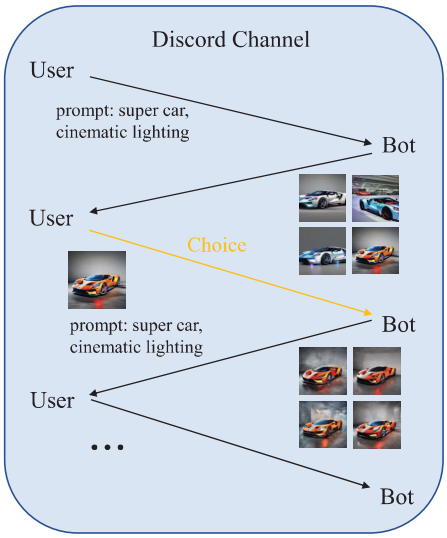


Figure 4. Interactions in the Discord channel. The human choice is highlighted in orange.

interaction follows a pre-defined grammar, which allows us to extract human choice and related images using simple pattern-matching techniques.

Data format and statistics. Finally, we obtain a total of 98,807 images generated from 25,205 prompts. Each prompt corresponds to several images, among which one image is chosen by the user as the preferred one, while others are non-preferred negatives. Each prompt corresponds with a varying number of images. 23,722 prompts have four images, 953 prompts have three images and 530 prompts have two images. The number of images for each prompt

	IS	FID
Preferred	16.27 ± 0.56	38.2
Non-preferred	16.23 ± 0.53	37.7

Table 1. The IS and FID of both preferred images and non-preferred images in our collected dataset. The FID is computed by comparing against a subset of images from the LAION-5B dataset that corresponds to the text inputs.

depends on the user's specifications in the generation request. Notably, the dataset exhibits a high level of diversity, with images generated across a broad range of themes. The dataset consists of choices made by 2,659 different users, and each user contributes at most 267 choices. Examples of the collected dataset can be found in Fig. 3. For further details on the dataset, we refer the readers to Fig. 9 in the appendix.

Privacy and NSFW contents. We observe that a small portion of images is generated with image condition (the condition image may be either generated or uploaded by the user). Since user-uploaded images may contain sensitive information or privacy, we do not include them in our dataset. For the images with potential NSFW content, we use the channel bot's NSFW detector to filter them out.

In this work, we utilize this dataset to study the existing metrics' correlation with human preferences, which will be introduced in Sec. 4. The dataset also serves as the training data for our human preference classifier, which is to be introduced in Sec. 5.

4. Existing Metrics

In this section, we show that the current mainstream evaluation metrics are not well correlated with human preferences on our dataset.

4.1. Metrics by Inception Net

Inception Score (IS) [36] and Fréchet inception distance (FID) [13] are two popular metrics used to evaluate the quality of generated images. Both of them perceive an image through an Inception Net [43] trained on ImageNet [34]. In this section, we investigate their correlation with human choices.

Inception Score (IS) measures the quality of generated images by computing the expected KL-divergence between the marginal class distribution over all generated images and the conditional distribution for a particular generated image, using the class probability predicted by the Inception Net. This metric is expected to capture both the fidelity and diversity of generated images. To determine the correlation between IS and human preferences, we compute IS for both the set of preferred and non-preferred images in our dataset. For each setting, we divide 20,000 images into 10 splits and reported the mean and standard deviation of IS computed on them. Our results, as shown in Tab. 1, indicate no significant difference between the preferred and non-preferred images.

Fréchet Inception Distance (FID) measures the similarity between the embedding feature of generated and real images. This is achieved by fitting the embedding features into a multivariate Gaussian distribution and computing their Fréchet distance. To define the target distribution, FID requires a set of real images. However, in the case of images generated from user-provided prompts, such as in our dataset, the target distribution is defined by users' intention, which can only be inferred from text prompts. To address this, we randomly sample 10,000 text prompts from our dataset, and for each prompt, we query the LAION [38] dataset via the official api to find the closest image, which is taken as a "pseudo ground truth" for that prompt. This provides a set of real images aligned with the users' intentions. We randomly sample 10,000 images from both the preferred and non-preferred split of the collected dataset to compute FID with the real images. Our results, as shown in Tab. 1, reveal no significant difference between the preferred and non-preferred images in terms of FID. This suggests that FID may not be a reliable metric for evaluating human preference.

Discussion. IS and FID may suffer from the following three issues when evaluating human preference. Firstly, generated images often contain shape artifacts, as shown in Fig. 1. However, classification-based CNNs tend to be biased towards image texture rather than shape [10], making them be likely to ignore shape artifacts in generated images. Additionally, the domain gap can pose a problem. While the evaluation model is trained on real images from ImageNet [34], the generated images in our dataset exhibit a wide range of styles and themes, from oil painting portraits to digital art of cyborgs. As a result, the ImageNet-trained model may not have meaningful representations for these diverse images [2]. Furthermore, these metrics are limited by their single-modal nature, which means that they cannot infer user intentions by accessing prompts, unless the target images are known or provided as we do.

4.2. Metrics by CLIP

Thanks to the large and diverse set of training data, CLIP is better at encoding images from various domains compared to ImageNet-trained models. Moreover, it can capture users' intentions by encoding text prompts, making it a plausible choice for evaluating the alignment between a prompt and a generated image [26, 30, 3, 32]. Aesthetic Score Predictor [37] is another CLIP-based tool for image quality evaluation, which has been utilized to filter the training data for Stable Diffusion [32]. In this section, we evaluate the capability of these tools in predicting human

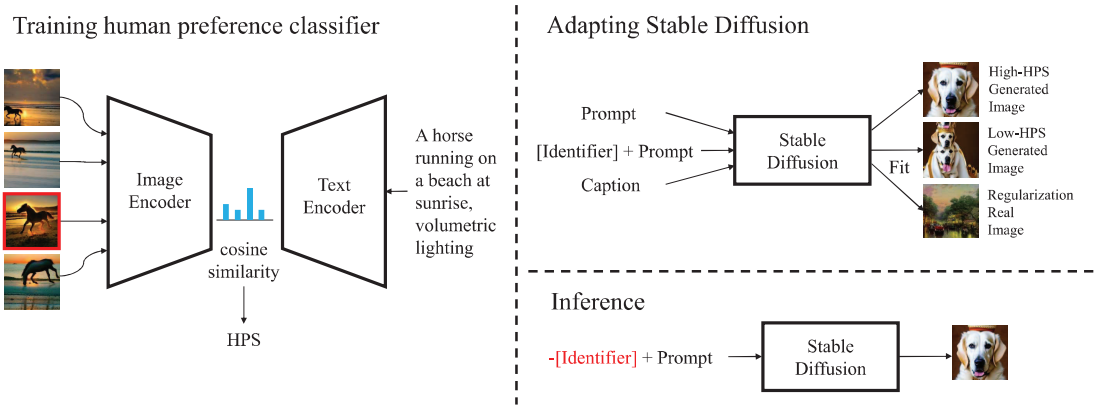


Figure 5. Left: training human preference classifier to derive HPS. Right: adapting Stable Diffusion to generate preferable images. During training, the Stable Diffusion is tuned to associate the concept of non-prefer with the prompt prefix [Identifier]. During inference, [Identifier] is used as the negative prompt in classifier-free guidance.

Model	Text	Preference Acc. (%)
Random guess	✗	26.1
Human	✓	42.0
CLIP ViT-L/14	✓	32.9
CLIP RN50x64	✓	33.1
Aesthetic Classifier	✗	31.4
HP classifier (ours)	✓	43.5

Table 2. Preference Acc. refers to the human choice prediction accuracy on 5,000 user choices. The Aesthetic Classifier makes prediction without seeing the text prompt.

choices, which is done by counting the accuracy of the human choice prediction task conducted on a split of 5,000 samples from our dataset.

CLIP score is derived as the cosine similarity between the prompt embedding and the image embedding computed by CLIP. We evaluated the performance of ViT-L/14 and RN50x64 models, which are the largest open-source CLIP models for transformer and CNN architecture. Our results, presented in Tab. 2, demonstrate that both CLIP models exhibit superior performance over random guessing. However, we will show in Sec. 7.1 that the CLIP score does not correlate well with human choices. Nevertheless, we will also show that it can be further fine-tuned on our dataset to better align with human preferences.

Aesthetic score is based on a pre-trained ViT-L/14 CLIP image encoder, which is adapted to the task of aesthetic score prediction by adding a MLP layer on top of the CLIP image encoder. The MLP is trained on several aesthetic datasets, including both real images and generated images (e.g., AVA [25], SAC [28]) to predict aesthetic scores ranging from 1 to 10. Unlike CLIP, the aesthetic classifier does not condition on the prompt, so the image with the highest predicted score is taken as the model choice. As shown in Tab. 2, the aesthetic classifier also exhibits better-than-chance accuracy in predicting user choice, indicating the importance of the aesthetic aspect of an image in human decision-making.

5. Human Preference Score

We first train a human preference classifier to predict the human choice based on the prompt, and then derive HPS based on the trained classifier.

Human preference classifier We fine-tune the ViT-L/14 version of CLIP on our dataset to better align with human preferences. Each sample in the training set contains one prompt along with $n \in \{2, 3, 4\}$ images, among which only one image is preferred by the user. The model is trained to maximize the similarity between the embedding of the text prompt computed by the CLIP text encoder and the embedding of the preferred image computed by the CLIP visual encoder, while minimizing the similarity for non-preferred images. By fine-tuning on human choices of generated images, the model is encouraged to better align with human preferences.

Human preference score (HPS) is derived from the human preference classifier. We define HPS as:

$$\mathrm{HPS}(\mathrm{img}, \mathrm{txt}) = 100 \cos(enc_v(\mathrm{img}), enc_t(\mathrm{txt})),$$

where enc_v and enc_t are the visual encoder and the text encoder of the human preference classifier. We multiply the cosine similarity by a factor of 100 for better visualization.

6. Better Aligning Stable Diffusion with Human Preferences

HPS can be used to guide diffusion-based generative models to better align with human users. We argue that the

misalignment between generated images and human preferences is a problem of missing "awareness" rather than model capacity. To address this issue, we propose to adapt the generative model by explicitly distinguishing preferred images from non-preferred ones. Our solution is straightforward and intuitive. We construct another dataset consisting of prompts and their newly generated images, which we categorize as either preferred or non-preferred using our previously trained human preference classifier. For the non-preferred images, we modify their corresponding prompts by prepending a special prefix. By adapting Stable Diffusion on this dataset via LoRA [17], we enhance the model's ability to learn the concept of non-preferred images, which can subsequently be avoided during inference.

Constructing training data. We construct the training data from the "`large_first_1m`" split of DiffusionDB [45], and a subset of the pre-train dataset of Stable Diffusion (LAION-5B) for regularization. DiffusionDB [45] is a large-scale dataset of generated images along with their text prompts. For images from DiffusionDB, we first compute HPS for each image-prompt pair. After that, we group the images by their prompts, and for each prompt T, we add the image I^* with the highest HPS into the training data if it passes the following criteria:

$$p > \frac{\alpha}{n},$$

where n is the number of images with the same prompt, and α is a hyper-parameter that controls the selectivity. p is given by:

$$p = \frac{\exp(HPS(I^*, T))}{\sum_{I \in B} \exp(HPS(I, T))},$$

where B is the set of images with the same prompt. Similarly, we construct the non-preferred subset by the same criteria, but using negative HPS. Finally, we get a mixed dataset of generated images and real images, where the non-preferred generated images are identified by their prompt prefix.

Adapting Stable Diffusion. We adopt LoRA [17] to adapt Stable Diffusion to the training data, in which the parameters of the original model are kept frozen, and the {key, query, value, out} projection matrices are augmented with a low-rank residual. LoRA does not add new parameters to the model, since the learned projection matrices can be merged into the base model once trained. During training, we use the prompt as the caption for generated images. For non-preferred images, we prepend a special identifier before each of their captions (we choose "Weird image." as the special identifier in our case). During inference, the special identifier is used as the negative prompt for classifier-free guidance [14] to avoid generating non-preferred images.

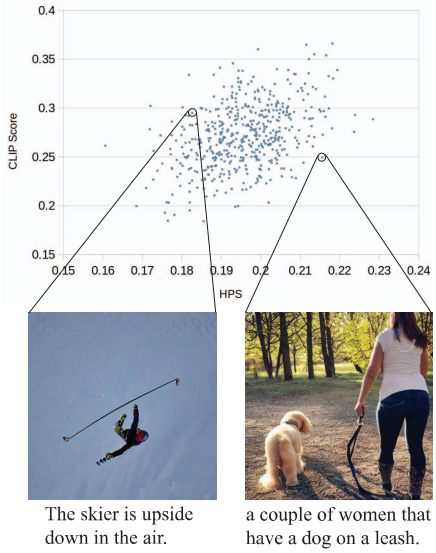


Figure 6. Correlation between HPS and CLIP score. While the CLIP score emphasizes more on the direct matching between the image content and the text prompt, HPS emphasizes more on the aesthetic quality of images.

7. Experiments

In this section, we firstly validate the reliability of HPS in Sec. 7.1, and then in Sec. 7.2, we introduce our experiments of adapting Stable Diffusion.

7.1. HPS

Implementation details of human preference classifier. We use 20,205 samples from our dataset during training, which contains 20,205 prompts and 79,167 images. We use the ViT-L/14 version of CLIP in our experiments. We fine-tune the last 10 layers of the CLIP image encoder and the last 6 layers of the text encoder. The model is trained by the AdamW optimizer [19] with a learning rate of 1.7×10^{-5} for 1 epoch. The batch size is 5. The learning rate decays with a cosine learning rate schedule. Weight decay is set as 3.1×10^{-3}. Instead of using the original data augmentation of random resized crop, we directly resize the longest edge of the image to 224, and then pad zeros to make the shorter edge increase to 224. We empirically find that fixing the aspect ratio of the image is beneficial. The hyper-parameters are tuned via Bayesian optimization.

Alignment with human. As shown in Tab. 2, the trained model significantly outperforms CLIP in the human choice prediction task. Due to the strong diversity of human preferences, the accuracy is even higher than our human participants.

Generalization. We evaluate HPS' generalization capability towards other generative models by user studies. In this experiment, we let the human preference classifier and

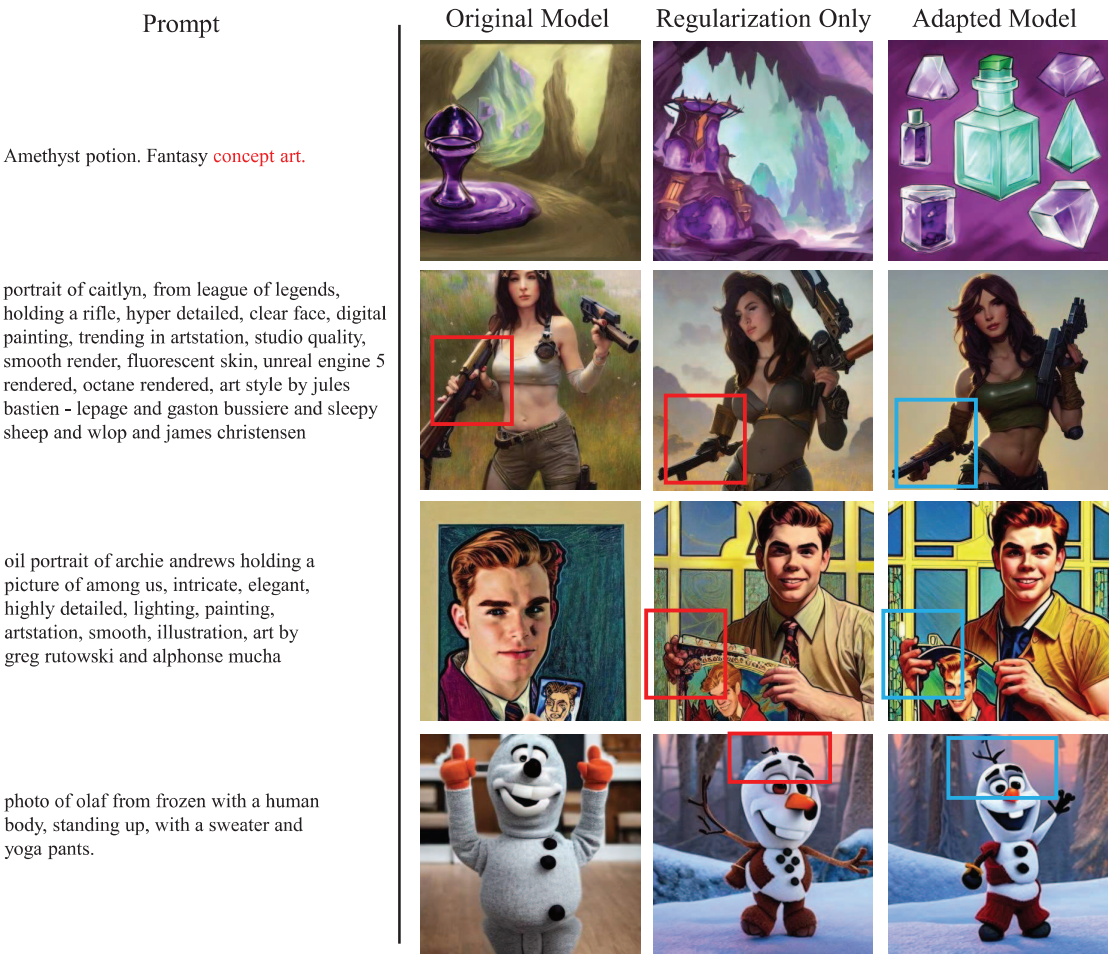


Figure 7. Comparison of images generated by the original model, the regularization-only model, and our adapted model. "Regularization Only" refers to a head-to-head setting against the "Adapted Model", where Stable Diffusion is adapted without HPS-labeled images. Images in the same row are generated with the same prompt and random seed. The prompts are sampled from DiffusionDB. The adapted model can better capture the user intention from the prompt, and generate more preferable images with fewer artifacts.

	Agreement (%)
Human vs. Human	63.5 ± 4.3
CLIP vs. Human	56.8 ± 1.7
HPS vs. Human	61.5 ± 1.1

Table 3. Agreement on comparing images generated by Stable Diffusion and DALL·E.

several human participants evaluate 398 pairs of images. In each pair, the images are generated by DALL·E [30] and Stable Diffusion [32] with the same text prompt. The prompts are randomly sampled from DiffusionDB [45], which is a large database of images and prompts sourced from the Stable Foundation Discord channel. We filter out the NSFW prompts by the indicator provided in DiffusionDB [45].

In Tab. 3, we evaluate the agreement between the predictions from humans, CLIP, and HPS. The agreement is computed by averaging the similarity of the prediction of each participant. HPS is better aligned with human preference compared to CLIP score, and its agreement with humans is close to the agreement between humans. It shows that HPS can generalize toward images generated by other models. We refer the readers to the supplementary material for a full list of images and choices made in this user study.

Correlation with CLIP score. In Fig. 6, we visualize the correlation between HPS and CLIP score. The text prompts are randomly sampled from the COCO Captions [5] dataset, and the images are generated by Stable Diffusion [32]. We can see that HPS has a positive correlation with CLIP score, but emphasizes more on the aesthetic quality of an image. However, HPS put less importance on the direct matching between image contents and text prompts, which can be interpreted as a visual analogy of "alignment tax" introduced

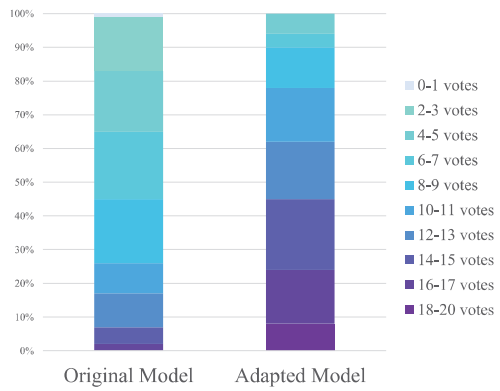


Figure 8. Human evaluation results on images generated from 100 randomly sampled prompts. The color represents the number of positive votes received from 20 participants.

in [27].

7.2. Better Aligning Stable Diffusion with Human Preferences

Implementation details. We use the Stable Diffusion [32] v1.4 for all our experiments. α is set to 2.0 for both preferred images and non-preferred images when constructing the training set. The constructed training set contains 37,572 preferred generated images and 21,108 non-preferred generated images. The regularization images are from a 625k subset of LAION-5B filtered by the aesthetic score predictor with a threshold of 6.5. 200,231 regularization images participate in training. We only fine-tune the UNet of Stable Diffusion, while keeping the VAE and the text encoder frozen during training. The rank is set to 32 in LoRA [17]. The LoRA weights are trained for 10k iterations with the AdamW [19] optimizer with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-2}, which is kept constant during training. We use a batch size of 40 in our experiments. For inference, we run the diffusion process by 50 steps for each image with PNDM [22] noise scheduler. We use the default guidance scale of 7.5 for classifier-free guidance [14].

Human evaluation. We compare our trained model with the original Stable Diffusion by conducting user studies. In this study, we randomly sample 100 user-provided prompts from DiffusionDB [45]. For each prompt, we generate an image from both models with the same random seed for fair comparison, resulting in 100 pairs of generated images for the user study. We ask 20 participants to read the prompt, and then choose between the image generated by our trained model and the original Stable Diffusion based on their preference. In Fig. 3, we visualize our result by showing the percentages of images with different numbers of positive votes. The adapted model significantly outperforms the original model. 74% of the images generated by the adapted model has more than 10 votes, while the number is 22% for the original model. A screenshot of the user-study interface is presented in Fig. 12 in the appendix.

	FID ↓	Aesthetic Score [37] ↑	CLIP Score [29] ↑	HPS ↑
SD 1.4	19.72	5.90	0.2816	0.1898
Adapted model	19.35	6.06	0.2831	0.1916

Table 4. Comparison between the original SD v1.4 and the adapted model.

Qualitative Evaluation. In Fig 7, we show some typical cases of improvement. We compare the original model, the regularization-only model, and the adapted model. The adapted model is trained with both real regularization images and generated images with HPS preference labels. The regularization-only model is a head-to-head comparison with the adapted model, which is trained by removing the generated images from the training set and is trained exclusively on regularization images for the same number of steps. The results show that the adapted model can better capture the user intention from the prompt, as shown in the first row. The last three rows show that training with generated images mitigates the problem of unnatural limbs. We refer the readers to Fig. 7 and Fig. 10 in the appendix for more examples.

Quantitative Evaluation. In Tab. 4, we compare the adapted model with the baseline on FID, Aesthetic Score, CLIP Score and HPS. The FID [13] is computed on 10k images from the LAION [38] dataset. CLIP Score [29] and HPS are computed on prompts from DiffusionDB [45].

8. Limitations

There are several limitations about the dataset. The collected dataset contains generated prompts and images of public figures. We choose to mark them out instead of removing them to keep the diversity of the dataset. Despite the diversity of the dataset, we are also aware that it only represents the preference of a small portion of people in the world, and it may be biased towards a certain group of people that are active in the Stable Foundation Discord channel. Another potential bias about this dataset is that a large portion of text prompts are written by experienced Stable Diffusion users. These prompts are very likely to be tweaked to activate the potential of Stable Diffusion and deviate from normal language habits.

9. Conclusion

In this work, we study human preferences on a large-scale dataset of generated images. We find that the previous evaluation metrics for generative models are not well aligned with human preferences, but the CLIP model can be fine-tuned into a human preference classifier to better

align with human choices. Then, we show a simple yet effective method to adapt the generative model to generate more preferable images with the guidance of human preference score. We hope our work can inspire the community to explore new possibilities of human-aligned AI research.

Acknowledgement

This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by General Research Fund of Hong Kong RGC Project 14204021. Hongsheng Li is a PI of CPII under the InnoHK. This project is also supported by SenseTime Collaborative Research Grant.

References

[1] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shin ichi Maeda. DQN-TAMER: Human-in-the-Loop Reinforcement Learning with Intractable Feedback. *ArXiv*, abs/1810.11748, 2018. 2

[2] Shane T. Barratt and Rishi Sharma. A Note on the Inception Score. *ArXiv*, abs/1801.01973, 2018. 4

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 2, 4

[4] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *ECCV*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, 2012. 2

[5] 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*, abs/1504.00325, 2015. 7

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 30, 2017. 2

[7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 2

[8] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, P. Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. *ArXiv*, abs/2212.05032, 2022. 2

[9] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. ERNIE-ViLG 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *CVPR*, pages 10135–10145, 2023. 2

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv*, abs/1811.12231, 2018. 2, 4

[11] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, pages 1462–1471. PMLR, 2015. 2

[12] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing Prompts for Text-to-Image Generation. *ArXiv*, abs/2212.09611, 2022. 2, 3

[13] 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 *NeurIPS*, 2017. 1, 2, 4, 8

[14] Jonathan Ho. Classifier-Free Diffusion Guidance. *ArXiv*, abs/2207.12598, 2022. 6, 8

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[16] Oleksii Holub. DiscordChatExporter. https://github.com/Tyrrrz/DiscordChatExporter, 2022. 3

[17] Edward Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021. 2, 6, 8

[18] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *ECCV*, pages 668–685. Springer, 2022. 2

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6, 8

[20] Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 119, 2022. 2

[21] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, P. Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning Text-to-Image Models using Human Feedback. *ArXiv*, abs/2302.12192, 2023. 3

[22] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo Numerical Methods for Diffusion Models on Manifolds. *ArXiv*, abs/2202.09778, 2022. 8

[23] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 2

[24] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 2

[25] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. *CVPR*, pages 2408–2415, 2012. 5

[26] 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 *ICML*, 2021. 1, 2, 4

[27] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022. 3, 8

[28] John David Pressman, Katherine Crowson, and Simulacra Captions Contributors. Simulacra Aesthetic Captions. Technical Report Version 1.0, Stability AI, 2022. 2, 5

[29] 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 *ICML*, 2021. 2, 8

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, abs/2204.06125, 2022. 1, 2, 4, 7

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 2

[32] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR*, pages 10674–10685, 2022. 1, 2, 4, 7, 8

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115:211–252, 2014. 2, 4

[35] 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. *NeurIPS*, 35:36479–36494, 2022. 1, 2

[36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016. 1, 2, 4

[37] Christoph Schuhmann. CLIP+MLP Aesthetic Score Predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022. 4, 8

[38] 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. *NeurIPS*, 35:25278–25294, 2022. 4, 8

[39] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023. 2

[40] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *CVPR*, pages 1599–1610, 2023. 2

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 2

[42] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32, 2019. 2

[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *CVPR*, pages 2818–2826, 2015. 4

[44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 2

[45] 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*, abs/2210.14896, 2022. 2, 6, 7, 8

[46] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. *ArXiv*, abs/2302.13848, 2023. 2

[47] Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-ViLG: Unified Generative Pre-training for Bidirectional Vision-Language Generation. *ArXiv*, abs/2112.15283, 2021. 2