Stable Preference: Redefining Training Paradigm of Human Preference Model for Text-to-image Synthesis

Hanting Li[1⋆], Hongjing Niu[1], and Feng Zhao[1†]

MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei, 230026, China
{ab828658,sasori}@mail.ustc.edu.cn
{fzhao956}@ustc.edu.cn

Abstract. In recent years, deep generative models have developed rapidly and can generate high-quality images based on input texts. Assessing the quality of synthetic images in a way consistent with human preferences is critical for both generative model evaluation and preferred image selection. Previous works aligned models with human preferences by training scoring models on image pairs with preference annotations. These carefully annotated image pairs well describe human preferences for choosing images. However, current training paradigm of these preference models is to directly maximize the preferred image score while minimizing the non-preferred image score in each image pair through cross-entropy loss. This simple and naive training paradigm mainly has two problems: 1) For image pairs of similar quality, it is unreasonable to blindly minimize the score of non-preferred images and can easily lead to overfitting. 2) The human robustness to small visual perturbations is not taken into account, resulting in the final model being unable to make stable choices. Therefore, we propose Stable Preference to redefine the training paradigm of human preference model and a anti-interference loss to improve the robustness to visual disturbances. Our method achieves state-of-the-art performance on two popular text-to-image human preference datasets. Extensive ablation studies and visualizations demonstrate the rationality and effectiveness of our method.

Keywords: Human preference model · Training paradigm · Synthetic image quality assessment

1 Introduction

Recent text-to-image generation methods have enabled the generation of high-quality photorealistic images from a piece of text. In particular, auto-regressive [3,5,28] and diffusion-based methods [21,27,31] have shown advantages in diversity and stability. To evaluate their performance, Inception Score (IS) [34] and

⋆ This work was done when Hanting Li was an intern at Huawei Noah's Ark Lab.
†The corresponding author is Feng Zhao.

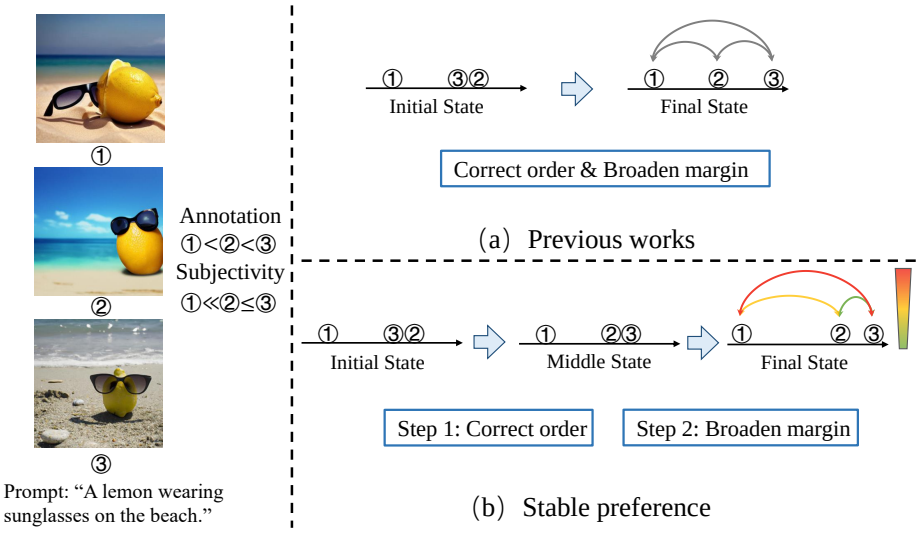


Fig. 1: Comparison between previous training paradigm and our stable preference. (a) Previous methods simultaneously performed order correction and broaden margin and treat every image pair equally (refer to Sec. 3.1)). (b) Stable preference first corrects the preference order, and then mainly expands the margin between images with significant difference (refer to Sec. 3.3).

Fréchet Inception Distance (FID) [8] remain the most popular metrics. Since these models synthesize various images by taken different random noise as input, the quality of the generated images may vary greatly. Therefore, measuring the quality of a single image is equally important as evaluating the model performance, while FID and IS are not suitable for evaluate a single image. CLIP [25] aligns text and images in the embedding space by pre-training on large-scale image-text pairs and can be used to measure the semantic relevance of between images and text. However, the human preference is influenced by not only the image-text alignment but also by other factors such as fidelity and aesthetic.

To facilitate research in this area, many carefully annotated preference datasets are proposed. These datasets can be roughly divided into two types. The first type directly collects image pairs and corresponding texts, and then the annotators only need to make a choice in each pair of images based on the text [13]. The second type collects a image group (usually contain 4∼10 images) for each text, and the annotators are asked to rank each set of images according to human preference [37–39]. Although there are differences in the collection and annotation of these datasets, the training paradigms of the preference models are almost identical. First, training samples will be uniformly processed into image pairs. For the dataset consist of image groups, in a group containing K images, at most $\mathcal{C}_K^2$ different comparison pairs can be obtained. Then choice within each pair of images is analyzed as a binary classification issue. And a preference model can be trained using cross-entropy loss.

Though these meticulously annotated image pairs encapsulate human preferences for image selection, extant training paradigms fail to fully leverage them, which is predominantly exhibited in two aspects. *Firstly, the image selection process is not strictly dichotomous.* Indiscriminately minimizing the score of non-preferred images, particularly where similar quality is exhibited (e.g., ② and ③ in Figure 1), is irrational and heightens the risk of model overfitting. *Secondly, the current model displays sensitivity towards minute visual disturbances.* Contrastingly, human preferences remain largely unaltered despite these minor perturbations. As illustrated in Figure 2, humans do not change their choices because of these small visual perturbations [2]. Nevertheless, when there is a minimal quality disparity between images, these perturbations can potentially lead to alterations in the choices made by the preference models. To address the above issues, we devise stable preference (SP), a novel training paradigm for the human preference model (HPM). It initially prioritizes the alignment of preferences (order) and then mainly enlarges the margin between images possessing significant differences, as illustrated in Figure 1. Furthermore, we formulate an anti-interference loss aimed at mitigating the sensitivity of the preference model to minor visual perturbations. This methodical progression ensures a more robust training process, thereby effectively thwarting premature over-differentiation of image pairs with similar quality. In summary, our contributions are as follows:

- We designed an anti-interference loss to reduce the model sensitivity to small visual perturbations, which further align the preference model with human scoring behavior.
- We propose a novel training paradigm for human preference models, named stable preference, which reduces the risk of overfitting.
- We conduct extensive experiments on two popular human preference datasets. Stable preference achieves best performance and can serve as a standard paradigm for training preference models. The effectiveness of our method is further corroborated by the results from visualizations.

2 Related Work

2.1 Metrics for Text-to-image Generation

To gauge the efficacy of various text-to-image generative models, some metrics have been introduced to measure the quality of the generated images. Among them, Fréchet Inception Distance [8] and Inception Score [34] are still the most prevalently used metrics. FID quantifies the distributional disparity between synthetic and real-world images, while the IS measures the fidelity and diversity of the generated images. Nonetheless, both of them are single-modal evaluation metrics, focusing solely on the image aspect. Consequently, using them to measure the quality of images generated from text conditions is suboptimal. Unlike IS and FID, which can only provide a single evaluation score, Sajjadi et al.

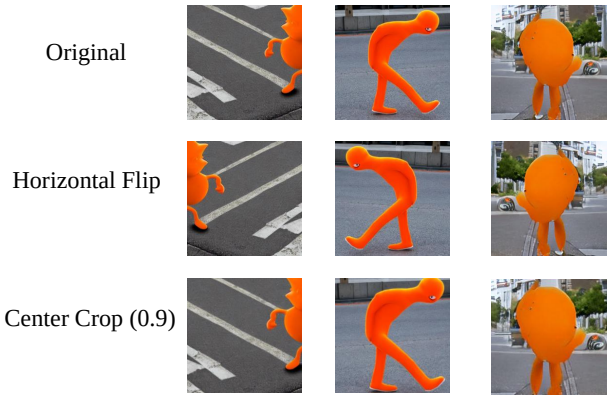


Fig. 2: Examples of small visual perturbations. These images are generated from "The image is of an anthropomorphic orange walking on a sidewalk". 0.9 is the side length ratio.

proposes to measure the distance between the generated and the reference distribution from two aspects (i.e., the precision and recall for distributions) [33]. Besides, there are also many other metrics designed to measure image diversity and fidelity [14, 20], while these metrics are not designed for conditional generative models.

To ascertain the congruity between generated images and user intent (i.e., input prompt), Xu et al. proposed R-Precision which utilized a text-image similarity model to calculate the top-1 retrieval accuracy from one hundred text candidates for each generated image [40]. Thanks to pre-training on 400 million image-text pairs [26], CLIP-R-Precision can be applied to evaluate a wider range of text-to-image generation models without extra training [23]. Compared with retrieval-based metrics, CLIP scores can directly measure the alignment of images and text. Another popular pipline is to measure the cyclic consistency between generated caption from the generated image and the input prompt [10]. Specially, the semantic object accuracy (SOA) is devised for evaluate each individual object through object detection models [9].

2.2 Human Preference Models

Human evaluation of images generated under text conditions is a complex process that involves an exhaustive evaluation of several factors such as text-image alignment, fidelity, and aesthetics. However, the aforementioned metrics are only able to address a fraction of these factors, thereby resulting in deviations from human preferences. To align with human preferences more accurately, the most direct approach involves the collection of a human preference dataset, following

which a human preference model can be trained on it. For example, the aesthetic predictors can score the aesthetic of each image by fine-tuning MLP layers that take the CLIP embedding as input on AVA [19] and LAION-aesthetic [35] to align with human aesthetic judgments. Some works directly exploit general multimodal models (e.g., GPT-4V and MiniGPT-4) and carefully design manual instructions to score each image on multiple aspects [42, 44]. Nonetheless, considering these models are not customized image evaluation models, their scoring quality is not uniformly reliable or stable, potentially leading to failure in more challenging scenarios.

In order to measure human preferences more comprehensively, some works manually rank image pairs or image groups, such as Pick-a-Pic [13], ImageReward [39], and human preference dataset [38]. Annotators are required to annotate image pairs or image groups based on detailed annotation document. Then the human preference models are trained on the image pairs annotated with human choice. These models can be trained via cross-entropy loss by transforming the problem into a binary classification task on each image pair. Since most preference models are fine-tuned from vision-language pre-training models, such as CLIP [25] and BLIP [15], these models usually converge quickly and are prone to overfitting. Existing methods take a variety of tricks to prevent overfitting, such as freeze some backbone layers [38] and adding dropout layers. However, these models still exhibits sensitivity to training hyperparameters [39]. Therefore, we propose stable preference to redefine the training paradigm, reducing the risk of overfitting and improving the stability of the human preference model.

3 Method

3.1 Problem Definition

Human preference models are designed to score on any text and image pairs. When the text is fixed, it can be used to evaluate the relative quality of any two images under the textual conditions. Let (I_1, I_2) refer to an image pairs generated from the same text T. A preference model F_θ is expected to score I_1 and I_2 under the condition T, which can be defined as,

$$\begin{aligned} S_1 &= \mathcal{F}_\theta(I_1, T) \\ S_2 &= \mathcal{F}_\theta(I_2, T). \end{aligned} \tag{1}$$

Then S_1 and S_2 can then be used to make choice between I_1 and I_2 (e.g., if $S_1 > S_2$, then I_1 is better than I_2 under the condition T). If the choice made by the preference model aligns with the human choice in the majority of circumstances, it can be used to automatically evaluate the quality of synthetic images and generative models. Therefore, the most direct training paradigm of human preference model is to regard the image selection problem as a binary classification problem, and the corresponding loss function can be formulated as follows,

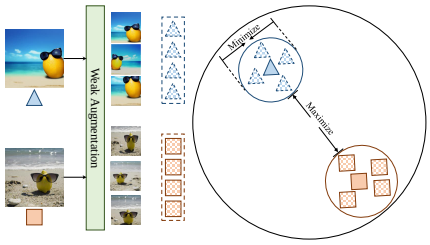


Fig. 3: Schematic of the anti-interference loss.

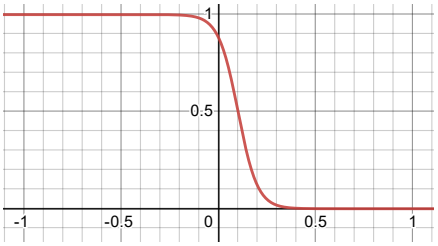


Fig. 4: The curve of $\frac{1}{1+e^{(\Delta S-0.1)/0.05}}$. ΔS is the independent variable of the function.

$$\mathcal{L}_{pref} = \sum_{i=1}^{2} y_i \log \hat{y}_i, \tag{2}$$

with

$$\hat{y}_i = \frac{\exp(\mathcal{F}_\theta(I_i, T))}{\sum_{j=1}^{2} \exp(\mathcal{F}_\theta(I_j, T))}. \tag{3}$$

Where $[y_1, y_2]$ is the preference label. Specifically, $[y_1, y_2]$ takes a value of $[1, 0]$ if I_1 is preferred, $[0, 1]$ if I_2 is preferred. This training paradigm has been widely adopted in the training of most previous human preference models [13, 37, 39].

3.2 Anti-interference Loss

Human usually focus more on the central region of an image [22] and maintain their preference in the presence of small perturbations [2]. Because of this characteristic, they usually do not change their choices due to small visual disturbances. However, the preference model cannot learn this human evaluation habit well under the existing training paradigm. To fix this gap, we proposed anti-interference loss to lessen the sensitivity of the preference model to small visual perturbations. Specifically, as shown in Figure 3, for image pair (I_1, I_2) with text T and preference label $[y_1, y_2] = [1, 0]$ (i.e., I_1 is better than I_2), we perform k independent weak data augmentation on each image and obtain $2k$ extra views with small visual perturbations $(\{I_1^i\}, \{I_2^i\}), \quad i \in \{1, 2, 3, ..., k\}$. The corresponding $2k$ preference scores can be obtained through the preference model and are recorded as,

$$S_j^i = \mathcal{F}_\theta(I_j^i, T) \qquad\qquad i \in \{1, 2, 3, ..., k\}. \tag{4}$$

As shown in Figure 3, the goal is to minimize the difference between different views of the same image while maximizing the margin between two images. Then the anti-interference loss can be formulated as,

$$\mathcal{L}_{ai} = -\log \frac{e^{dist_{inter}}}{e^{dist_{inter}} + e^{dist_{intra}}} \tag{5}$$

with

$$\begin{aligned} dist_{inter} &= \min(\{S_1^i\} \cup S_1) - \max(\{S_2^i\} \cup S_2), \\ dist_{intra} &= \frac{\sum_{j=1}^{2} \max(\{S_j^i\} \cup S_j) - \min(\{S_j^i\} \cup S_j)}{2}, \\ & \qquad i \in \{1, 2, 3, ..., k\}. \end{aligned} \tag{6}$$

Where $dist_{inter}$ and $dist_{intra}$ are inter-image distance between I_1 and I_2 and the mean intra-image distance between different augmented views of a same image, respectively. In contrastive learning, strong augmentation is often used to generate different views so that the network learns consistent representations. $\mathcal{L}_{ai}$ is designed to make the preference model have stable scoring ability under small perturbations, which is why we use weak augmentation.

3.3 Stable Preference

While the aforementioned training paradigm is straightforward and intuitive, the human selection process is not as binary as characterized by the cross-entropy function. Accordingly, we propose a two-step training paradigm named Stable Preference.

Step 1: Correct the preference order. The main task of the preference model is to select the image according to human preferences, rather than maximizing the score difference in each image pair. Given this, in the initial phase, the preference model should primarily learn from the image pairs in which it made incorrect selections and correct them. Besides, as shown in Figure 1, we should only ensure that the model can just make choice correctly, and not further expand the distance between images (i.e., $0 < S_1 - S_2 < \epsilon$ when I_1 is better, where ϵ is a small positive value). Therefore, the loss function of this stage is,

$$\mathcal{L}_1 = \frac{\mathcal{L}_{pref} + \mathcal{L}_{ai}}{1 + e^{(\Delta S - b)/\tau}}, \tag{7}$$

where ΔS is the difference between the preference score of the better image and that of the worse image (i.e., if I_1 is better than I_2, then $\Delta S = S_1 - S_2$). b and τ are two hyperparameters that dictate the minimal discernible difference in scores between images. More specifically, by setting the appropriate b and τ, when $\Delta S < 0$ (i,e, the preference model does not make the correct choice), $\mathcal{L}_1 \approx \mathcal{L}_{pref} + \mathcal{L}_{ai}$, and when $\Delta S > b + \tau \ln(\frac{1-\epsilon}{\epsilon})$, $\mathcal{L}_1 < \epsilon(\mathcal{L}_{pref} + \mathcal{L}_{ai})$. More intuitively, we show the function curve of $\frac{1}{1+e^{(\Delta S - b)/\tau}}$ when setting b and τ to 0.1 and 0.05 in Figure 4. When the model is able to make the correct choice, $\mathcal{L}_1$ will quickly decay to near zero, which ensures that the model focuses on correcting the preference order.

Step 2: Broaden the margin After learning the correct preference order, our objective shifts towards enlarging the score discrepancy between images whose differences are pronounced, which prevent the over-differentiation of image pairs assessed as having similar quality in the first step. The loss of the j-th image pair in a mini batch can be written as follows,

$$\mathcal{L}_2 = \frac{e^{\Delta S_j}}{\sum_{i=1}^{N} e^{\Delta S_i}}(\mathcal{L}_{pref} + \mathcal{L}_{ai}), \tag{8}$$

where N is the mini-batch size, and ΔS_i stands for the score difference of the i-th image pair in each mini batch.

4 Experiments

To verify the effectiveness of our method, we conduct extensive experiments on ImageReward [39] and human preference dataset v2 [37]. The datasets and implementation details are described below.

4.1 Dataset

ImageReward Datasets [39] utilizes a diverse selection of real user prompts from DiffusionDB [36]. After filtering out similar prompts, the dataset finally yields 10,000 candidate prompts, each accompanied by 4 to 9 sampled images from DiffusionDB. The annotators are asked to evaluate the image according to the annotated document in terms of image-text alignment, fidelity, and harmlessness. After rating for each aspects, annotators will finally rank each image group. The dataset finally contain valid annotations for 8,878 prompts, resulting in 136,892 compared pairs. Among them, 8,000 prompts are used for training, 466 prompts are for testing, and the rest 412 prompts are for validation.

Human Preference Datasets v2 (HPD v2) [37] is the latest human preference dataset for preference model training, preference model testing, and benchmarking generative models. In order to obtain more diverse prompts, HPD v2 collect 108k prompts from both COCO captions [1] and DiffusionDB [36]. Specially, the prompts from DiffusionDB are cleaned by ChatGPT. Since the training split is not yet open source, we only use the test split and image benchmark of the HPD v2 to verify the cross-dataset generalization performance and benchmarking the text-to-image generative model. The test split consists of 400 groups of images, which corresponds to 400 individual prompts. Each group contains 9 images generated from 9 popular text-to-image generative models and is annotated by 10 distinct annotators to ensure the annotation quality. The image benchmark part contains 3,200 prompts for four styles, including "Animation", "Concept-art", "Painting", and "Photo". Various prompts ensure that the model can generate diverse images to fully evaluate the performance of the generative model.

DrawBench [32] collect a comprehensive and challenging set of prompts that support the evaluation and comparison of text-to-image models. In total, DrawBench contains 200 different prompts.

Table 1: Comparison of human preference models sensitivity to small visual perturbations on HPD v2 and ImageReward datasets. "ORG" represents the baseline result on original test split. "HP" and "CC" stand for horizontal flip and center crop, respectively. Numbers in brackets represent the side length ratio of the center crop. SP represents our stable preference training paradigm.

Method	Dataset	ORG	HP&CC (0.97)	HP&CC (0.95)	HP&CC (0.93)	HP&CC (0.90)
HPS v2	HPD v2	83.3	82.2 (-1.1)	82.2 (-1.1)	81.8 (-1.5)	81.7 (-1.6)
ImageReward		74.2	73.7 (-0.5)	73.6 (-0.6)	73.6 (-0.6)	74.0 (-0.2)
SP (CLIP-L)		77.2	77.3 (+0.1)	77.0 (-0.2)	76.9 (-0.3)	77.0 (-0.2)
SP (CLIP-H)		80.7	81.4 (+0.7)	80.3 (+0.4)	80.4 (+0.3)	80.7 (+0.0)
HPS v2	ImageReward	65.7	64.8 (-0.9)	63.8 (-1.9)	64.2 (-1.5)	63.9 (-1.8)
ImageReward		65.2	64.5 (-0.7)	64.8 (-0.4)	64.8 (-0.4)	65.3 (+0.1)
SP (CLIP-L)		66.3	65.7 (-0.6)	65.6 (-0.7)	65.9 (-0.4)	66.0 (-0.3)
SP (CLIP-H)		66.8	67.4 (+0.6)	66.4 (-0.4)	66.5 (-0.3)	66.7 (-0.1)

4.2 Implementation Details

We fine-tune CLIP-H and CLIP-L [11] on ImageReward for 30,000 steps with the AdamW optimizer [18] following a cosine learning rate schedule, with 3,000 steps for optimizing with $\mathcal{L}_1$ and the rest for optimizing with $\mathcal{L}_2$. The initial learning rate is set at 2×10^{-6}. The AdamW optimizer is used with a mini-batch size of 64 and the weight decay is 0.2. A warm-up period of 1,500 steps is adopted. As in previous work [38, 39], we freeze the parameters of shallow layers in CLIP to prevent overfitting. Specifically, we train the last 18 layers of the CLIP image encoder and the last 8 layers of the CLIP text encoder. b and τ are is set at 0.3 and 0.2 by default. To calculate the anti-interference loss, random horizontal flipping and random resized cropping (the lower and upper bounds for the random area of the crop is $[0.85, 1]$) are adopted as the weak augmentation to obtain 3 extra views for each image. The selection accuracy of all image pairs on the test split is taken as the evaluation metric. All the experiments are conducted on two NVIDIA V100 GPUs with PyTorch toolbox [24].

4.3 Quantitative Results

Sensitivity to Small Visual Perturbations: To verify the robustness of our method to small visual perturbations, we conducted experiments on HPD v2 and ImageReward datasets. We adopt horizontal flipping and center cropping to introduce weak visual perturbations. For center cropping, we conducted experiments on several side length ratio settings that can hardly change human preference. To provide an intuitive understanding, we suggest referring to Figure 2 which illustrates the maximum perturbation under a side length ratio of 0.9. The results in Table 1 show that the performance of our method is more stable than other latest methods under different perturbations, which suggests that introducing anti-interference loss make the preference model less sensitive to small visual perturbations. Specially, under some perturbation settings our

Table 2: Evaluation of the anti-interference loss and our stable preference training paradigm.

Settings		Model	ImageReward
L_{ai}	Stable Preference		
✗	✗	CLIP-L	63.1
✔	✗		64.0
✗	✔		65.8
✔	✔		**66.3**
✗	✗	CLIP-H	63.9
✔	✗		64.5
✗	✔		66.2
✔	✔		**66.8**

Table 3: Evaluation of different weight assignment methods in step 2.

Settings	HPD v2	ImageReward
$e^{-\Delta S_j} / \sum_{i=1}^{N} e^{-\Delta S_i}$	79.5	66.2
$\frac{1}{N}$	80.1	65.8
$e^{\Delta S_j} / \sum_{i=1}^{N} e^{\Delta S_i}$	**80.7**	**66.8**

method even surpasses the performance without perturbation. This may due to the random cropping used in anti-interference loss that weights the center area of the image slightly higher than the edge area, which is consistent with the human habit that tend to focus the center area of the image [12].

Effectiveness of anti-interference loss and stable preference training paradigm. We evaluate the effectiveness of anti-interference loss and stable preference training paradigms on ImageReward dataset. The results in Table 2 show that both designs can boost model performance. Among them, the improvement of SP is relatively significant. The main reason is that the anti-interference loss is mainly designed to enhance the robustness of the model to small visual perturbations.

Whether we should focus on image pairs with larger ΔS? In step 2, we mainly focus on image pairs with pronounced difference. But in some other tasks, image pairs with small differences are given more attention [16]. Therefore, we conduct experiments to demonstrate that focusing on image pairs with larger ΔS is beneficial for subjective understanding task such as image preference. In Table 3, we compared giving similar image pairs a higher weight and giving all image pairs the same weight with our stable preference. The results show that the more conservative learning strategy we use can achieve better performance.

Table 4: Comparison with state-of-the art methods on test split of ImageReward dataset. † CLIP-H is initialized with the HPS v2 checkpoint.

Method	ImageReward
CLIP-L [11, 25]	54.8
CLIP-H [11, 25]	57.1
Aesthetic Score Predictor [35]	57.4
HPS v1 [38]	61.2
PickScore [13]	62.9
ImageReward [39]	65.1
HPS v2 [37]	65.7
Single Human vs. Single Human	65.3
Single Human vs. Averaged Human	53.9
Stable Preference (CLIP-L)	66.3
Stable Preference (CLIP-H)	66.8
Stable Preference (CLIP-H†)	**68.0**

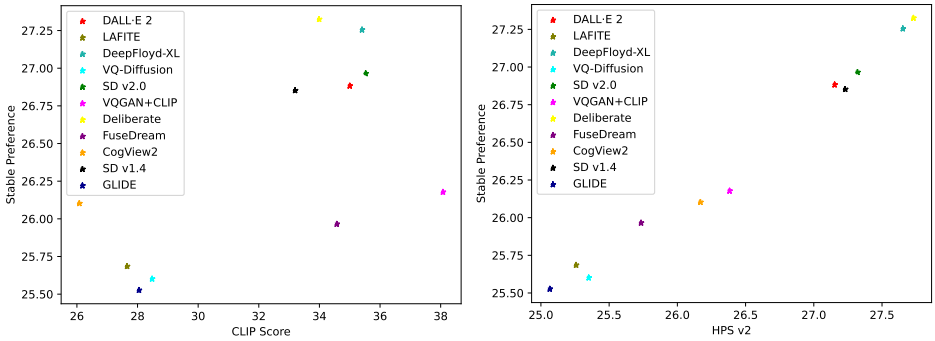
Comparison with State-of-the-Arts. We compare current state-of-the-art human preference models on the test split of ImageReward dataset. The results in Table 4 show that fine-tuning both CLIP-H and CLIP-L models under the SP paradigm can outperform previous methods. In particular, we also fine-tune the CLIP-H model initialized by the checkpoint of HPS v2 and obtained the highest performance of 68.0.

Evaluation of cross-domain performance. Although we achieved state-of-the-art performance compared with previous models, this comparison is not strictly fair because most of the methods in Table 4 are trained on different datasets (e.g., LAION-aesthetic [35], HPD v2 [37], and Pick-a-Pic [13]). Besides, these models are fine-tuned from similar vision-language models, such as CLIP [11] and BLIP [15]. Therefore the quality of their dataset is one of the biggest factors affecting their final performance. Therefore, in Table 5, we uniformly train all models on the training split of ImageReward and verify their performance on the test set of HPD v2. Under such a setting, we can not only fairly verify the effectiveness of our training paradigm, but can also facilitating an evaluation of cross-domain generalizability. The results show that the stable preference significantly exceeds the results of training through cross-entropy loss, once again validating the superiority of stable preference.

Benchmarking Latest Text-to-image Generative Models by Stable Preference. In addition to automatically selecting preferred images, another major use of human preference models is to evaluate the performance of text-to-image generative models. In order to make it convenient for future work to use stable preference (CLIP-H†) to verify the effectiveness of their own text-to-image generative models, we evaluated the latest models in academia and industry on the DrawBench [32] and HPD v2 benchmarks in Table 6. We follow the protocol

Table 5: Comparison of cross-domain performance. All the models are trained on the training set of ImageReward and tested on the test split of HPD v2. † CLIP-H is initialized with the HPS v2 checkpoint.

Method	HPD v2
CLIP-L [11, 25]	72.8
CLIP-H [11, 25]	74.8
BLIP [15]	74.2
Single Human vs. Single Human	78.1
Single Human vs. Averaged Human	85.0
Stable Preference (CLIP-L)	77.2
Stable Preference (CLIP-H)	80.7
Stable Preference (CLIP-H†)	**82.5**



(a) Stable preference & CLIP Score **(b)** Stable preference & HPS v2

Fig. 5: Correlation between stable preference and other human preference models. The model score is calculated by the average score of all images in DrawBench [32].

in [37] to report the mean and standard deviation of 10 groups of images on HPD v2.

4.4 Visualization

Correlation between Stable Preference and Other Human Preference Models. In Figure 5, we show the correlation between stable preference and other preference models in benchmarking text-to-image generative models. These results indicate that our method is positively correlated with existing human scoring models, but can better align with the human preferences according to the quantitative results in Tables4/5.

Image Selection Based on Human Preference Models. We generate 100 candidate images for each text through Stable Diffusion v1.4 [30], and show the top-1 choice of each model in Figure 6. It can be seen that our stable preference

Table 6: Evaluation of recent text-to-image generative models through our stable preference (CLIP-H†). For DrawBench, we report the average score of all images. For HPD v2, we divide 800 prompts of each style into ten groups of 80, and report the mean and standard deviation of 10 groups, which is consistent with [37].

Model	HPD v2				DrawBench [32]
	Photo	Concept-art	Animation	Painting	
GLIDE [21]	25.03±0.245	24.12±0.111	24.25±0.151	24.27±0.150	25.53
VQ-Diffusion [7]	25.82±0.189	25.24±0.118	25.45±0.156	25.50±0.108	25.60
LAFITE [43]	26.16±0.179	25.37±0.049	25.48±0.090	25.38±0.132	25.67
FuseDream [17]	25.79±0.221	25.54±0.071	25.61±0.073	25.44±0.163	25.97
CogView2 [4]	26.29±0.227	26.66±0.092	26.58±0.106	26.51±0.091	26.10
Latent Diffusion [29]	26.03±0.158	25.72±0.091	26.10±0.102	25.84±0.156	26.16
VQGAN + CLIP [6]	25.90±0.185	26.15±0.077	26.13±0.121	26.04±0.116	26.18
DALL·E mini	26.16±0.206	25.94±0.114	26.32±0.103	25.97±0.101	26.33
Versatile Diffusion [41]	26.59±0.202	26.07±0.121	26.40±0.124	26.21±0.108	26.44
Stable Diffusion v1.4 [30]	26.85±0.193	26.57±0.083	27.03±0.145	26.63±0.131	26.85
DALL·E 2 [27]	26.80±0.153	26.59±0.089	27.26±0.117	26.68±0.173	26.88
Epic Diffusion	26.98±0.152	26.79±0.073	27.27±0.141	26.89±0.098	26.91
Stable Diffusion v2.0 [30]	27.01±0.185	26.83±0.063	27.29±0.146	26.82±0.144	26.97
Openjourney	27.02±0.148	27.00±0.059	27.48±0.132	27.07±0.115	27.00
ChilloutMix	27.15±0.162	27.13±0.048	27.72±0.127	27.18±0.159	27.06
MajicMix Realistic	27.19±0.180	27.21±0.058	27.67±0.171	27.25±0.148	27.13
Dreamlike Photoreal 2.0	27.28±0.204	27.31±0.074	27.79±0.127	27.32±0.115	27.25
DeepFloyd-XL	27.16±0.147	26.91±0.080	27.49±0.084	26.90±0.120	27.26
Realistic Vision	27.24±0.184	27.38±0.068	27.93±0.110	27.44±0.123	27.31
Deliberate	27.23±0.172	27.40±0.061	27.92±0.127	27.39±0.101	27.33

has good performance in evaluating image-text alignment, fidelity, and aesthetic. For the third prompt, the image selected by our method not only consider the concepts appearing in the text, i.e., girl, sunflower and hallway, but also take into account the reasonable arrangement of these concepts in the image.

5 Limitation & Future Works

This work mainly focuses on the improvement of the visual side, but text and vision are equally important for training human preference model. We summarize several potential research directions not covered in this work to provide hints for future work.

- Preference models need to be able to adapt to various forms of user prompts, and should have consistent scoring criteria for prompts that have the same semantics but different expressions.
- Evaluating image quality based on the complete prompts is the most coarse-grained setting. Designing a finer-grained preference model will help increase the interpretability of the preference model.
- Given the subjectivity of the task, decision-making based on a singular preference model may exhibit bias. Reasonable integration of different preference models should be able to further approximate the average human preference.

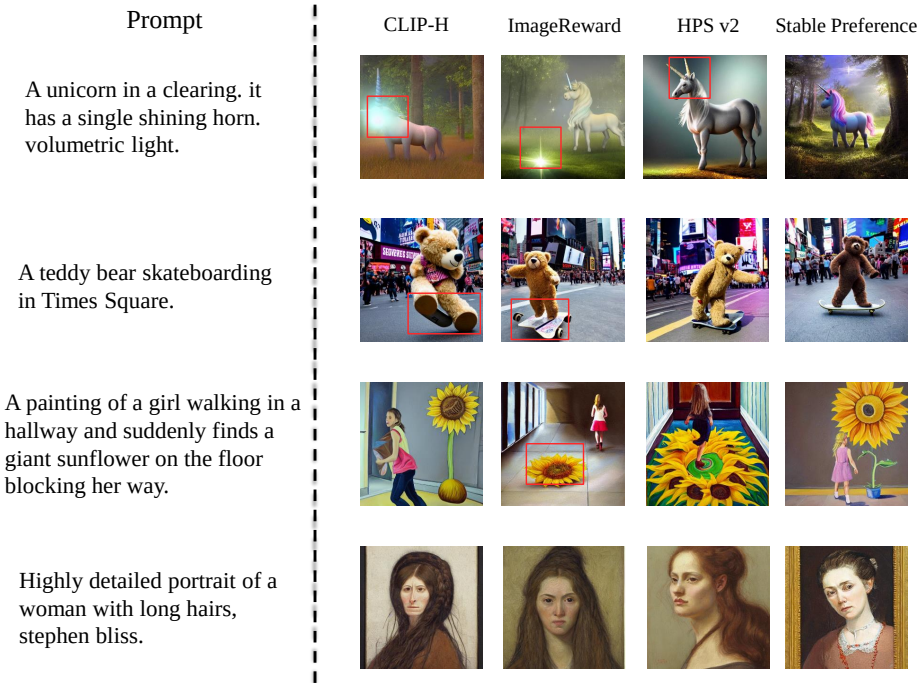


Fig. 6: Top-1 images out of 100 (Stable Diffusion v1.4) generations selected by stable preference and other human preference models.

We anticipate that the aforementioned topics will generate significant interest in the field of preference modeling. As we contemplate potential solutions to these issues, there is room to further optimize the training paradigm of preference models. This aspect will be our prime focus in our future work.

6 Conclusion

In this work we propose Stable Preference, a new paradigm for human preference models. Training in the order of first aligning preference order and then mainly broaden the margin between images with significant difference effectively mitigates the risk of overfitting. Besides, we designed an anti-interference loss to reduce the sensitivity of preference model to small visual perturbations that do not affect human preferences, which further improves the robustness of the model. In terms of experiments, we conducted extensive experiments to verify the effectiveness of the proposed method, and provided a benchmark for evaluating text-to-image generative models based on SP model. In addition, we also provide visualization results to prove that SP model is closer to human preferences than existing models.

Acknowledgments

This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

1. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
2. Digirolamo, G.J., Hintzman, D.L.: First impressions are lasting impressions: A primacy effect in memory for repetitions. Psychonomic Bulletin & Review **4**(1), 121–124 (1997)
3. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Adv. Neural Inf. Process. Syst. **34**, 19822–19835 (2021)
4. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. Adv. Neural Inf. Process. Syst. **35**, 16890–16902 (2022)
5. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 12873–12883 (2021)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 12873–12883 (2021)
7. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 10696–10706 (2022)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. **30** (2017)
9. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1552–1565 (2020)
10. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 7986–7994 (2018)
11. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). `https://doi.org/10.5281/zenodo.5143773`, `https://doi.org/10.5281/zenodo.5143773`, if you use this software, please cite it as below.
12. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 2106–2113. IEEE (2009)
13. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. arXiv preprint arXiv:2305.01569 (2023)
14. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Adv. Neural Inf. Process. Syst. **32** (2019)

15. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proc. Int. Conf. Mach. Learn. pp. 12888–12900. PMLR (2022)
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 2980–2988 (2017)
17. Liu, X., Gong, C., Wu, L., Zhang, S., Su, H., Liu, Q.: Fusedream: Training-free text-to-image generation with improved CLIP+GAN space optimization. arXiv preprint arXiv:2112.01573 (2021)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 2408–2415. IEEE (2012)
20. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: Proc. Int. Conf. Mach. Learn. pp. 7176–7185. PMLR (2020)
21. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
22. Palmer, S.E., et al.: Aesthetic issues in spatial composition: Effects of position and direction on framing single objects. Spatial vision **21**(3), 421–450 (2008)
23. Park, D.H., Azadi, S., Liu, X., Darrell, T., Rohrbach, A.: Benchmark for compositional text-to-image synthesis. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. **32** (2019)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
27. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
28. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proc. Int. Conf. Mach. Learn. pp. 8821–8831. PMLR (2021)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 10684–10695 (2022)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 10684–10695 (2022)
31. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. **35**, 36479–36494 (2022)

32. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. **35**, 36479–36494 (2022)
33. Sajjadi, M.S., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing generative models via precision and recall. Adv. Neural Inf. Process. Syst. **31** (2018)
34. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Adv. Neural Inf. Process. Syst. **29** (2016)
35. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Adv. Neural Inf. Process. Syst. **35**, 25278–25294 (2022)
36. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.14896 (2022)
37. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
38. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 2096–2105 (2023)
39. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. arXiv preprint arXiv:2304.05977 (2023)
40. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 1316–1324 (2018)
41. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 7754–7765 (2023)
42. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v (ision) as a generalist evaluator for vision-language tasks. arXiv preprint arXiv:2311.01361 (2023)
43. Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., Sun, T.: LAFITE: Towards language-free training for text-to-image generation. arxiv 2021. arXiv preprint arXiv:2111.13792
44. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)