Learning Multi-dimensional Human Preference for Text-to-Image Generation

Sixian Zhang*, Bohan Wang*, Junqiang Wu*, Yan Li†, Tingting Gao, Di Zhang, Zhongyuan Wang
Kuaishou Technology

Abstract

Current metrics for text-to-image models typically rely on statistical metrics which inadequately represent the real preference of humans. Although recent work attempts to learn these preferences via human annotated images, they reduce the rich tapestry of human preference to a single overall score. However, the preference results vary when humans evaluate images with different aspects. Therefore, to learn the multi-dimensional human preferences, we propose the Multi-dimensional Preference Score (MPS), the first multi-dimensional preference scoring model for the evaluation of text-to-image models. The MPS introduces the preference condition module upon CLIP model to learn these diverse preferences. It is trained based on our Multi-dimensional Human Preference (MHP) Dataset, which comprises 918,315 human preference choices across four dimensions (i.e., aesthetics, semantic alignment, detail quality and overall assessment) on 607,541 images. The images are generated by a wide range of latest text-to-image models. The MPS outperforms existing scoring methods across 3 datasets in 4 dimensions, enabling it a promising metric for evaluating and improving text-to-image generation. The model and dataset will be made publicly available to facilitate future research. Project page: https://wangbohan97.github.io/MPS/.

1. Introduction

"There are a thousand Hamlets in a thousand people's eyes."

–Vissarion Belinsky

Text-to-image generative models [10, 14, 15] have achieved remarkable advancements in recent years, and these models have the capability to generate high-fidelity and contextually relevant images based on textual descriptions (i.e., prompts). To evaluate the quality of generated images, several evaluation metrics are proposed, including Inception Score (IS) [16], Fréchet Inception Distance (FID) [5], and CLIP Score [12]. However, these statistical metrics fall short of aligning with human perceptual preferences. For instance, metrics like the IS or the FID, although indicative of image quality to some extent, might not necessarily reflect how a human observer would rate the image in terms of fidelity, coherence, or aesthetic appeal.

Figure 1. As humans evaluate images from different perspectives, their preference for the images also varies. Specifically, when examining the images in the top row, the image on the left stands out in terms of aesthetic appeal, though it falls short in semantic alignment (e.g., two boats on the river) compared to its counterpart on the right. In the case of the bottom row, both images are aesthetically pleasing, yet the image on the right is marred by poor detail quality (e.g., as signified by the red bounding boxes around the distorted hand and foot).

Contrary to these statistical metrics, several approaches [6, 20–22, 24] turn towards human-centric evaluations, where generated images are manually annotated according to human preferences. Subsequently, models are trained with these annotations to predict preference scores. However, these approaches typically utilize a single score to summarize all human preferences, overlooking the multi-dimensionality of human preferences. As Fig. 1 shows, the preference results differ when humans evaluate images from various perspectives. Therefore, using a single-dimensional evaluation method is insufficient in capturing the broad

*These authors contributed equally to this work.
†Corresponding author.

Table 1. **Comparisons of text-to-image models quality databases**. Our Multi-dimensional Human Preference (MHP) dataset achieves significant advancements over existing work in three aspects, including prompt collection, image generation, and preference annotation. Moreover, it constitutes the largest dataset both in generated images and preference annotations. Note that the Diffusion DB only contains generated images but lacks annotations of human preferences. Besides, KOLORS is an internal dataset derived from in-house platform for designer, which provides ∼10w prompts.

Dataset	Prompt collection		Image Generation		Preference annotation	
	Source	Annotation	Source	Number	Rating	Dimension
DiffusionDB [19]	DiffusionDB	×	Diffusion (1)	1,819,808	0	None
AGIQA-1K [24]	DiffusionDB	×	Diffusion (2)	1,080	23,760	Overall
PickScore [6]	Web Application	×	Diffusion (3)	583,747	583,747	Overall
ImageReward [22]	DiffusionDB	×	Auto Regressive; Diffusion (6)	136,892	410,676	Overall
HPS [21]	DiffusionDB	×	Diffusion (1)	98,807	98,807	Overall
HPS v2 [20]	DiffusionDB, COCO	✓	GAN; Auto Regressive; Diffusion, COCO (9)	430,060	798,090	Overall
AGIQA-3K [7]	DiffusionDB	×	GAN; Auto Regressive; Diffusion (6)	2,982	125,244	Overall; Alignment
MHP	DiffusionDB, PromptHero, KOLORS, GPT4	✓	GAN; Auto Regressive; Diffusion (9)	607,541	918,315	Aesthetics, Detail, Alignment, Overall

range of personalized needs and preferences. To ensure a comprehensive evaluation of text-to-image synthesis outputs, it is crucial to learn and utilize multi-dimensional human preferences.

To learn the multi-dimensional human preferences, we propose the Multi-dimensional Human Preference (MHP) dataset. Compared to prior efforts [6, 20–22], the MHP dataset offers significant enhancements in prompts collection, image generation, and preference annotation. For the prompt collection, previous work [6, 21, 22] directly utilizes existing open-source datasets (e.g., Diffusion DB [19]) or datasets collated from the internet [6], overlooking the potential data bias of long-tail distribution. To this end, based on the categories schema of Parti [23], we annotate the collected prompts into 7 category labels (e.g., characters, scenes, objects, animals, etc.). For the underrepresented tail categories, we employ Large Language Models (LLMs) (e.g., GPT-4 [9]) to generate additional prompts. This process results in a balanced prompt collection across various categories, which is used for later image generation. For image generation, following previous work [7, 20], we not only utilize existing open-source Diffusion models and their variants, but also employ GANs and auto-regressive models to generate images. Consequently, we generate a dataset of 607,541 images, which are further used to create 918,315 pairwise comparisons of images for preference annotation. The quantity of image data constitutes the largest dataset of its kind. For the annotation of human preferences, contrary to the single annotation of existing work [6, 20, 21, 24], we consider a broader range of dimensions for human preferences and employ human annotators to label each image pair across four dimensions, including aesthetics, detail quality, semantic alignment, and overall score.

To learn human preferences, existing methods employ the pre-trained vision-language models (e.g., CLIP [11], BLIP [8]) to extract features from images and prompts independently, followed by computing the similarity between them. These methods then fine-tune the network utilizing the collected preference data. For learning the multi-dimensional preferences, a straightforward strategy is to train separate models for different preferences. However, such a simple strategy requires data re-collection and model re-training for the new preference. Moreover, due to the potential bias in single-preference data, a model trained under one preference condition often exhibits diminished performance when evaluated against other preferences. Therefore, we propose the Multi-dimensional Preference Score (MPS), a unified model capable of predicting scores under various preference conditions. Specifically, a certain preference is denoted by a series of descriptive words. For instance, the 'aesthetic' condition is decomposed into words such as 'light', 'color', and 'clarity' to describe the attributes of this condition. These attribute words are used to compute similarities with the prompt, resulting in a similarity matrix that reflects the correspondence between words in the prompt and the specified condition. On the other hand, features from images and text are extracted using a pre-trained vision-language model. Subsequently, two modalities are fused through a multimodal cross-attention layer. The similarity matrix serves as a mask merged into the cross-attention layer, which ensures that the text only related to the condition is attended to by the visual modality. Then the fused features are used to predict the preference scores. We evaluate our MPS model on both the existing human preference datasets (i.e., ImageReward [22] and HPS v2 [20]) and our MHP dataset. The experimental results indicate that our MPS model surpasses existing benchmarks in evaluating both overall and multi-dimensional prefer-

ences, establishing a new state-of-the-art in comparison with related work.

In summary, our main contributions are as follows:

- We introduce the Multi-dimensional Human Preference (MHP) datasets for evaluating text-to-image models. The MHP contains balanced prompts and the largest collection of images with multi-dimensional annotations. Based on MHP, we propose a standard test benchmark to evaluate existing text-to-image synthesis models.
- We propose the MPS model, which learns multi-dimensional human preferences and evaluates the scores of generated images under different preference conditions.
- Our MPS exhibits superior performance compared to existing methods across three datasets in predicting the overall preferences and multi-dimensional preferences.

2. Related work

2.1. Text-to-image Generation and Evaluation

The text-to-image task aims to synthesize realistic images from natural language description (i.e., prompt). Several work attempts to tackle this problem, including GANs [4, 25], auto-regressive [2, 3, 23] and diffusion models [10, 14, 15]. Among the previously mentioned methods, diffusion models gain significant attention for their exceptional performance. These methods are principally divided into two categories: latent-based and pixel-based approaches. The Latent Diffusion Model (LDM) [14] is notable as the first to introduce a latent-based diffusion model, leveraging an auto-encoder to map images into a latent space where the diffusion process is executed. Following this, Stable Diffusion has notably propelled the field forward by open-sourcing SD series [10]. In contrast, DALL·E 2 [13] and Imagen [15] are predicated on pixel-based diffusion models. Besides, the Imagen [15] integrates the large language model T5 XXL to achieve a text-to-image super-resolution diffusion model capable of producing highly realistic images. Current text-to-image models excel in creating high-quality images but often miss aligning with human preferences in real-world applications. For evaluating text-to-image models, several evaluation metrics are proposed to evaluate the quality of generated images, including Inception Score (IS) [16], Fréchet Inception Distance (FID) [5], and CLIP Score [12]. However, these statistical metrics fall short of aligning with human perceptual preferences. Our MPS provides a comprehensive evaluation for text-to-image generations, facilitating the evaluation of alignment with multi-dimensional human preference.

2.2. Learning human preferences

Currently, several studies attempt to collect and learn human preferences for the evaluation of text-to-image generation. They use the collected data to fine-tune visual-language models (VLMs) to align with human selections.

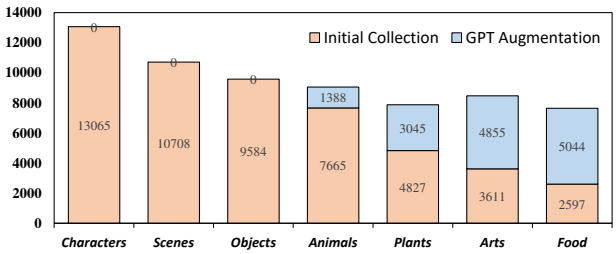

Figure 2. **Prompt collection**. The initially collected prompts exhibit a long-tail distribution across various categories. After prompt augmentation with GPT, we obtain a relatively balanced prompt dataset, which contains 66,389 prompts.

HPS [21] introduces the HPD dataset of human preference choices. The images are generated solely using the Stable Diffusion model with prompts from Diffusion DB. They train HPS model utilizing the human preference annotations of HPD to align it with human preference. Subsequently, they fine-tune the Stable Diffusion model under the guidance of HPS, leading to better generated images that are more preferred by human users. However, HPS is limited to a single generation model and a relatively small number of images. Furthermore, HPS v2 [20] introduces a larger dataset, employing 8 generative models, including Diffusion models, GANs, and Auto regressive models, and also incorporating captions from the COCO dataset. However, both HPS and HPS v2 primarily focus on overall human preferences, not considering the diversity of human tastes.

PickScore [6] proposes a web application designed to collect prompts and human preference annotations from real users. Unlike previous methods adopting prompts from existing datasets (e.g. DiffusionDB), the prompts of PickScore are directly generated by actual users. The dataset of PickScore is sizeable, however, it focuses only on the overall preferences, lacking detailed annotations for multi-dimension preferences.

ImageReward [22] employs four types of Diffusion models along with an auto-regression-based model. Their annotation of generated images is more detailed with scoring ranging from 0 to 7. Beyond overall satisfaction, they also consider annotations for alignment and fidelity. However, they merge alignment and fidelity into a single overall score, inadequately capturing the multi-dimensional dimensions of human preferences.

AGIQA-1k [24] and AGIQA-3k [7] utilize a range of generative models, including Diffusion, GAN, and Auto regressive models, to produce images. They consider both overall and alignment preferences. However, their dataset size is considerably smaller compared to existing work.

Our MHP dataset represents an advancement over previ-

Table 2. **Image generation**. The image sources of the MHP dataset consist of the images generated from 9 text-to-image generative models. Note that KOLORS is an internal model for in-house designer platform.

Source	Type	Images	Split
KOLORS	Diffusion	211,707	Train and test
DeepFloyd IF	Diffusion	27,311	Train and test
Stable Diffusion XL	Diffusion	89,176	Train and test
Openjourney v4	Diffusion	133,875	Train and test
Stable Diffusion v2.0	Diffusion	84,590	Train and test
Stable Diffusion v1.5	Diffusion	56,882	Train and test
VQGAN+CLIP	GAN	1,000	Test
LAFITE	GAN	1,000	Test
CogView2	Autoregressive	2,000	Test

ous work in terms of prompt collection, image generation, and preference annotation.

3. MHP Dataset

3.1. Prompt collection and annotation

Our prompts are carefully collected from several databases, including PromptHero[1], DiffusionDB[19] and KOLORS-dataset (an internal dataset derived from in-house platform for designers). Following the category schema from Parti[23], we further merge some categories, and finally determine 7 categories as illustrated in Fig. 2. The definitions and merging rules are detailed in the supplementary material. Based on these defined categories, we employ human annotators to label the initially collected 59,396 prompts. Additionally, the annotators are also required to filter out anomalous prompts, such as those that are incoherent, incomprehensible, or have punctuation errors. After annotation, we obtain 52,057 prompts. The distribution of these prompts, based on their categories, is depicted in Fig. 2. As the figure shows, the initially collected prompts exhibit a long-tail distribution across categories. Such category imbalances might lead to imbalanced generated images. Consequently, the human preferences learned from these imbalanced data could also be biased. As a result, we further expand our prompts.

By employing the GPT-4[9], we obtain additional prompts to supplement categories with initially low quantities (see supplementary materials for more examples of the generated prompts). These generated prompts are further refined by annotators to remove those incoherent or incomprehensible items. As shown in Figure 2, after supplementation, we obtain 66,389 prompts and the distribution of prompts across categories is balanced. These balanced prompts help us in learning more representative human preferences.

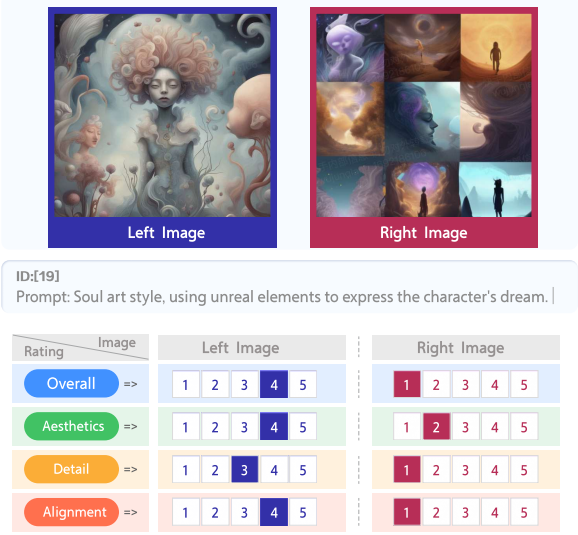

Figure 3. **Annotation interface**. Annotators are required to evaluate the preference for the given image pair on four dimensions, including aesthetics, detail quality (detail), semantic alignment (alignment) and overall score (overall)). Annotation scores are discrete values ranging from 1 to 5, which are subsequently normalized to Boolean values of 0 or 1. When the scores are tied, the normalized score is set to 0.5.

3.2. Image collection and annotation

As shown in Table 2, we utilize Diffusion models (such as the Stable Diffusion series and DeepFloyd IF), GANs, and AutoRegressive models to generate images based on the obtained prompts. Each model produces 2-4 images for every single prompt. The generated images come from a variety of model architectures, with various image resolution scales (e.g., 512×512, 1024×1024, 1366×768), and aspect ratios (e.g., 1:1, 16:9). This diversity ensures a comprehensive representation of the text-to-image models' generalization capability.

Images generated from the same prompt are paired together for comparison. To enhance the representativeness of these image pairs, the construction of image pairs sources not only from images generated by different models but also includes those produced by the same model using different random seeds. Based on these contrastive image pairs, we employ human annotators to evaluate the image pairs with our annotation interface as shown in Fig. 3. The annotators are required to evaluate the quality of the generated image pairs based on three sub-dimensions (i.e., aesthetics, text-image consistency, and detail) and one overall dimension (i.e., overall score). These four dimensions are defined as follows:

1. **Aesthetics**: annotators should measure the aesthetic quality of a generated image pair in terms of composition, light contrast, color matching, clarity, tone, style,

depth of field, atmosphere, and artistry of the image.

2. **Detail quality**: annotators should focus on the delicacy of image details such as texture, hair, and light and shadow, whether there is distortion in the face, hands, and limbs of the characters, whether there is a blurry overall view, object distortion, severe deformation.
3. **Semantic alignment**: annotators should evaluate the semantic consistency of the generated images with the prompts, and the evaluation includes measuring whether the generated image accurately matches the textual description (e.g., quantity, attributes, location, positional relationships) and whether there is missing or redundant content in the generated image.
4. **Overall assessment**: Based on the combination of above aspects and subjective preferences, the annotators assess the quality of each generated image from a holistic perspective.

The annotators rate all these scores of each image in the pair into five distinct levels (from 1 to 5), and the scores are eventually normalized to $[0, 1]$. The image annotation is completed by a crowdsourced team of 210 members. Before the official annotation, each member needs to perform pre-annotation, where any member whose annotation results have a high degree of inconsistency with that of the majority is disqualified. Eventually, 198 members participate in the annotation of generated image pairs, of which 170 members act as annotators and 28 members act as quality inspectors. Each image pair is annotated by three annotators respectively and the final result is averaged from these three annotation results. 20% of the annotated data is extracted and sent to the quality inspector for inspection. If there is significant difference in annotation results between annotators and inspectors, the annotated data is considered invalid and will be relabeled.

3.3. Statistics

In summary, we collect 66,389 prompts and employ 9 recent text-to-image models to generate 607,541 images. Based on these images, we construct 918,315 image pairs. Notably, 20% of these pairs are created using the same model but with different settings, while the other 80% are produced by different models, allowing for a wide range of comparisons. Each image pair is then annotated by 3 distinct annotators across 4 dimensions to enable the study of diversity in human preference.

We divide the annotated data into training, validation and test sets. The training and validation set contains 898,315 and 10,000 image pairs and the test set comprises 10,000 pairs. To ensure that the data distribution is representative, the test set includes not only images generated by Diffusion models but also those produced by GAN and Autoregressive models.

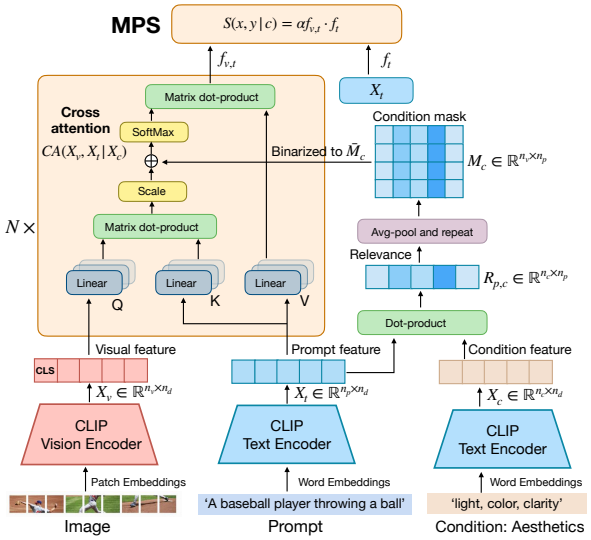

Figure 4. **The framework of Multi-dimensional Preference Score (MPS)**. The MPS takes the generated image, prompt and preference condition as the input, and predicts the quality (i.e. human preference) of the generated image under the given preference condition.

4. Multi-dimensional Preference Prediction

4.1. Model Structure of MPS

As shown in Fig. 4, we adopt CLIP [11] to initially extract features from images and prompts. Given a prompt x, a generated image y and a preference condition c, the visual feature of y is obtained by the vision encoder of the CLIP by $X_v = E_v(y)$, where $X_v \in \mathbb{R}^{n_v \times n_d}$. n_v is the token number of the image and n_d is the feature dimension. $X_t = E_t(x)$, where $X_t \in \mathbb{R}^{n_p \times n_d}$ and n_p is the token number of the text. $X_c = E_c(c)$, where $X_c \in \mathbb{R}^{n_c \times n_d}$ and n_p is the token number of word sets representing the preference. For the setup of word sets for each preference, please refer to Sec. 5.1. Previous works [6, 21, 22] typically utilize the first dimension of X_v and the last dimension of X_t to calculate the preference score, which loses a lot of detailed information of both image and text. Alternatively, we employ the full range of X_v and X_t and fuse two modalities through the Cross Attention (CA) module

$$CA\left(X_v, X_t\right) = \sigma\left(\frac{X_v W_q \left(X_t W_k\right)^T}{\sqrt{n_d}}\right) X_t W_v \quad (1)$$

where σ is the SoftMax activation function, and $W_* \in \mathbb{R}^{n_d \times n_d}$ are parameters and biases are omitted. Our motivation is computing the preference scores in different preference conditions. Specific words in the prompt should be given more attention based on different conditions, e.g., when aesthetics is considered as a condition, words in the prompt related to color, light, and clarity should be taken

Table 3. Main results of MPS and comparison methods on human preference evaluation. Preference accuracy (%) is calculated on ImageReward, HPD v2, and our MHP dataset.

ID	Preference Model	ImageReward	HPD v2	MHP (Overall)
1	CLIP score [11]	54.3	71.2	63.7
2	Aesthetic Score [17]	57.4	72.6	62.9
3	ImageReward [22]	65.1	70.6	67.5
4	HPS [21]	61.2	73.1	65.5
5	PickScore [6]	62.9	79.8	69.5
6	HPS v2 [20]	65.7	83.3	65.5
7	MPS (Ours)	**67.5**	**83.5**	**74.2**

into account more when calculating the score. Therefore, we propose a condition mask to highlight the relevant tokens while suppressing the irrelevant tokens. The condition is represented by a series of attribute words, e.g. the preference condition of 'Aesthetic' is represented by a set of words, including light, color and clarity. The relevance of prompt and condition is computed by $R_{p,c} = X_c X_t^T W_c + b_c$, where $R_{p,c} \in \mathbb{R}^{n_c \times n_p}$ and W_c and b_c are learnable parameters. The $R_{p,c}$ is averaged along the dimension n_c and then repeated n_v times to obtain the mask $M_c \in \mathbb{R}^{n_v \times n_p}$. The M_c is further binarized to $\bar{M}_c$, where elements below the similarity threshold are assigned to negative infinity and others are set to zero. Based on the condition mask $\bar{M}_c$, the Eq. 1 is further improved as

$$CA\left(X_v, X_t | X_c\right) = \sigma\left(\frac{X_v W_q \left(X_t W_k\right)^T}{\sqrt{d}} + \bar{M}_c\right) X_t W_v \tag{2}$$

The Eq. 2 ensures that the parts of the prompt that are relevant to the condition information receive more attention when computing the preference score. We adopt the first dimension (i.e. the cls token) of fused feature $CA\left(X_v, X_t | X_c\right)$ for further prediction, which is denoted as $f_{v,t}$ and $s \in \mathbb{R}^{1 \times n_d}$ Additionally, to prevent the issue of excessively short prompts leading to a scenario where no words of the prompt are related to the condition. In such a scenario, the cross-attention layer would be unable to capture the information from the prompt. Therefore, we supplement the $f_{v,t}$ with additional prompt feature f_t, where f_t is the last dimension of X_t. Consequently, the MPS is obtained by

$$S(x, y|c) = \alpha f_{v,t} \cdot f_t \tag{3}$$

where α is a learned scalar while x, y and c denote the prompt, image and preference condition, respectively.

4.2. Training

The input for our objective includes our scoring function MPS $S(x, y|c)$, a prompt x, two generated image y_1, y_2, a preference condition c and the preference score (annotated by human) p, where p takes a value of $[1, 0]$ for y_1 is preferred, $[0, 1]$ if y_2 is preferred, or $[0.5, 0.5]$ for ties. Following previous work [6], the training objective minimizes the KL-divergence between the annotation p and the softmax-normalized prediction

$$\begin{aligned} \hat{p}_{i,c} &= \frac{expS\left(x, y_i|c\right)}{\sum_{i=1}^{2} expS\left(x, y_i|c\right)} \\ L_P &= \sum_c \sum_{i=1}^{2} p_{i,c}(logp_{i,c} - log\hat{p}_{i,c}) \end{aligned} \tag{4}$$

We initialize the text and vision encoders, E_t and E_v, with parameters from the pre-trained CLIP-H model, while the remaining parameters are subject to random initialization. We train our MPS on MHP datasets for 30,000 steps, with a batch size of 128, a learning rate of 3e-6, and a warmup period of 500 steps.

5. Experiments

5.1. Experimental Setup

Preference condition setting. We utilize the following collection of word sets to represent human preferences: 1) Aesthetics: *light, color, clarity, tone, style, ambiance, artistry*; 2) Detail quality: *shape, face, hair, hands, limbs, structure, instance, texture*; 3) Semantic alignment: *quantity, attributes, position, number, location*; 4) Overall: *light, color, clarity, tone, style, ambiance, artistry, shape, face, hair, hands, limbs, structure, instance, texture, quantity, attributes, position, number, location*.

Evaluation setting. We select widely used statistical metrics for evaluating text-to-image models, namely the CLIP score [12] and Aesthetic score [17] for comparison. Additionally, we also choose methods that align with human preferences for evaluating text-to-image models, including Image Reward [22], HPS [20, 21] and PickScore [6]. Following previous works [6, 20], we utilize publicly available pre-trained models without finetuning for evaluation.

5.2. Evaluation Results

Overall Preference accuracy. Previous works on learning human preferences mostly focus on a singular, overall preference, i.e., summarizing human preferences with an overall score. For a fair comparison, we choose existing publicly available human preference datasets: the ImageReward test set [22] and the HPD v2 test set [20], along with our MHP dataset (using only data annotated with overall scores) to compare our method with relevant baselines. As shown in Tab. 3, our MPS demonstrates a better accuracy across these three datasets, indicating the strong generalization capability of our method.

Table 4. The evaluation of MPS and related scoring functions for the prediction of multi-dimensional human preferences(%).

ID	Preference Model	Overall	Aesthetics	Alignment	Detail
1	CLIP score [11]	63.67	68.14	82.69	61.71
2	Aesthetic Score [17]	62.85	82.85	69.36	60.34
3	ImageReward [22]	67.45	74.79	75.27	58.31
4	HPS [21]	65.51	73.86	73.86	62.05
5	PickScore [6]	69.52	70.95	70.92	56.74
6	HPS v2 [20]	65.51	73.86	73.87	62.06
7	MPS (Ours)	**74.24**	**83.86**	**83.87**	**85.18**

Multinational Preference accuracy. In addition to the overall score, we also compare the performance of previous works and our method in predicting multi-dimensional human preferences based on our MHP dataset. As Tab. 4 illustrates, the CLIP Score and Aesthetic score focus on specific types of preferences, only perform well in certain preferences (e.g., semantic alignment or aesthetics). However, they fall short in predicting other preferences compared to models trained on human preferences. Additionally, preference models [6, 20–22] generally perform better in overall score and some other dimensions but lack generalization in certain specific preferences (e.g., details)*. We illustrate the Fig. 5 to reveal the underlying reasons for this poor generalization. In the first and second rows of Fig. 5, the score functions exhibit a high correlation only with the trained preferences (e.g. semantic alignment and overall score), but perform poorly on other preferences. It is important to note that the prediction of different preferences is based on the same data, albeit with different preference annotations. This indicates that not all preferences are strongly correlated, which results in that improvements in one preference might come at the expense of others. Therefore, only learning a single score is inadequate in fully reflecting the complexity of human preferences. In contrast, our MPS learns human preferences with the condition mask from multiple dimensions and maintains high consistency across all dimensions of human preferences, as shown in the third row of Fig. 5. Besides, as Tab. 4 indicated, MPS outperforms the related works by a large margin in predicting the multi-dimensional human preference on the MHP dataset.

Visualization Results. Further, we aim to explore why our MPS exhibits strong generalization across various dimensions of human preferences, even some preferences (e.g., detail quantity) are not highly correlated with others. To this end, we visualized the attention map of images and prompts that MPS focuses on when predicting human preferences. As shown in Fig. 6, we employ GradCAM [18] and $f_{v,t}$ to generate attention heatmaps of the image, and utilize the values of M_c to represent the attention heatmap of the prompt. The visualization results indicate that our HPS attends to different regions of prompts and images depending on the specific preference condition. This is attributed to the condition mask, which allows only those words in the prompt related to the preference condition to be observed by the image. The condition mask ensures that the model predicts the preference with different inputs, and the model only needs to calculate the similarity between patches in the image and the retained partial prompt to determine the final score. Therefore, the selective focus enabled by the condition mask allows utilizing a unified model to predict multinational preferences effectively, even if some preferences have weak correlations with others.

*Since these methods have been trained solely to generate an overall score, we could only duplicate it for multinational preferences.

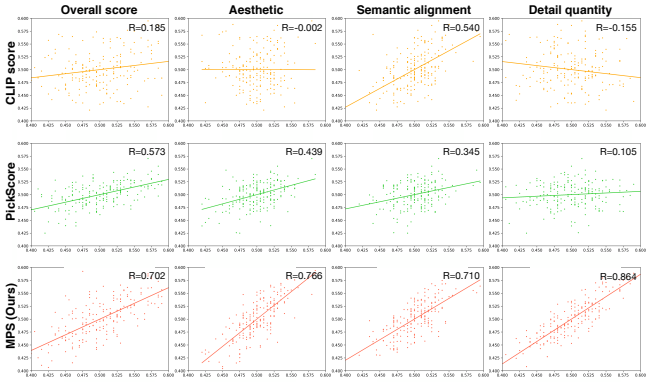

Figure 5. **Correlation between real user preferences and model predictions**. The x-axis of each subplot represents the annotated real human preferences, and the y-axis denotes the model's predictions. We examine three models: CLIP score, PickScore, and MPS (ours). Each subplot is annotated with the calculated correlation coefficient R-value, where a higher R-value indicates a closer alignment of the model's predictions with actual human preferences.

Ablation study. We conduct ablation studies to verify the effectiveness of each component, as illustrated in Tab. 5. Compared to the baseline† (i.e., PickScore), the cross-attention module enables more comprehensive integration between image and prompt, leading to improvement of prediction accuracy in overall score, aesthetics, and semantic alignment. However, the model still underperforms in detail quantity, which is less correlated with other preferences. The addition of the condition mask M_c alleviates this issue and improves the prediction performance across various preferences, especially in the detail quantity. Furthermore, we train separate models for different preferences.

†Note that the baseline shares the same network architecture with PickScore in Tab. 3 and 4, but is trained on our MHP dataset.

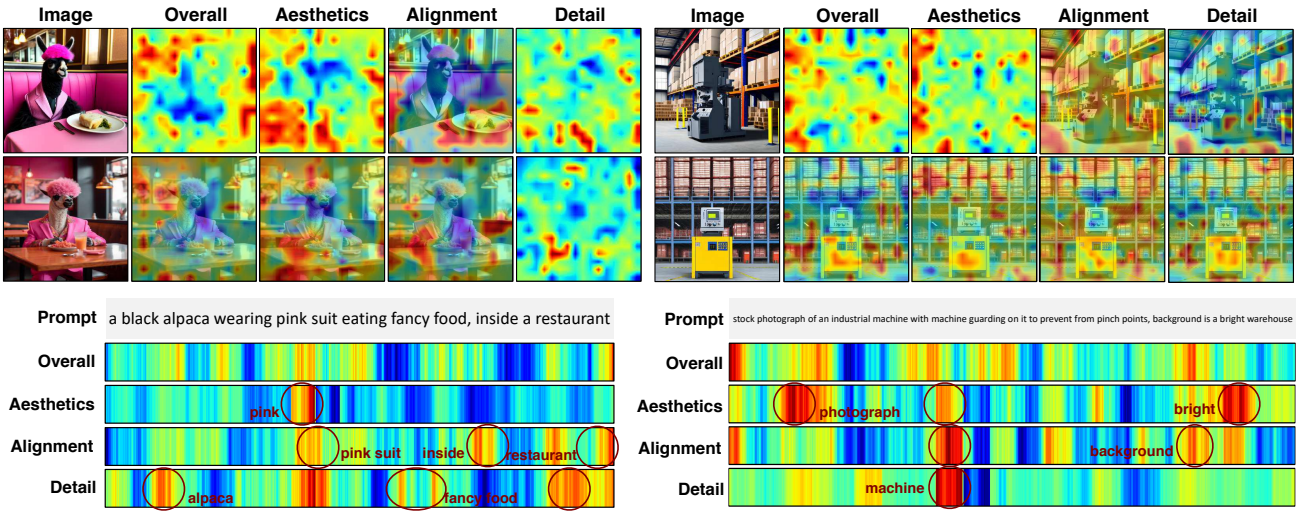

Figure 6. **Visualization**. We leverage Grad-CAM and the condition mask M_c to visualize attention heatmaps for the image and prompt. The condition mask results in the prompt focusing on words related to the preference condition. For Aesthetic, the model tends to focus on colors (e.g., *pink*) and lighting (e.g., *bright*). In the case of semantic alignment, the focus shifts to attributes (e.g., *pink suit*) and position (e.g., *inside*, *background*). For detail quantity, the model's attention is on instances (e.g., *alpaca*, *machine*, *fancy food*). On the image side, the model also tends to focus on areas of the image that correlate with the parts of the prompt receiving attention.

Table 5. **Ablation study for different modules used in MPS**. Base: PickScore is employed as the baseline model. CA: Cross-attention module. Mask: Preference condition mask M_c. Row 4 illustrates the models that have identical structures but are trained separately for each preference type.

ID	Module			Overall	Aesthetics	Alignment	Detail
	Base	CA	Mask				
1	✓			70.65	71.05	70.81	57.46
2	✓	✓		74.11	71.83	73.02	58.79
3	✓	✓	✓	**74.24**	**83.86**	**83.87**	**85.18**
4	Separately trained MPS			73.51	80.54	79.68	76.81

Experimental results indicate that models trained separately for each preference do not perform as well as those trained to learn multiple preferences simultaneously. We infer that more extensive annotations and unified training contribute to better model generalization. Ablation studies validate the effectiveness of each module, particularly the condition mask, in learning multiple preferences.

5.3. MPS Benchmark

Based on the MPS model and the collected MHP dataset, we introduce the MPS benchmark for evaluating text-to-image models across multiple dimensions. The MPS benchmark includes a set of evaluation prompts designed to assess the models on a total of 4,000 prompts, covering seven categories: characters, scenes, objects, animals, plants, arts, and food. Each category comprises 500 prompts. Our MPS assesses the images generated by the text-to-image models across four dimensions: Aesthetic, Semantic Alignment, Detail Quantity, and Overall Score. The scoring results can assist users in selecting superior models based on their personal preferences. Additionally, the scoring results can also enhance the generative models' performance by selecting more preferable images with higher MPS scores.

6. Conclusions

In this work, we introduce the Multi-dimensional Human Preference (MHP) dataset and Multi-dimensional Preference Score (MPS) to evaluate text-to-image models from multi-dimensional human preferences. The MHP dataset offers improvements over previous methods in prompt collection, image generation, and preference annotation, which comprises 918,315 human preference choices across four dimensions. These preferences include aesthetics, semantic alignment, detail quantity, and overall score. Additionally, to align the multi-dimensional human preferences, we propose the MPS, which employs a unified network to score generated images based on varying preference conditions. The MPS introduces a condition mask that retains words in the prompt related to the preference condition. Subsequently, the model integrates only the retained prompt with the image to compute the final score. MPS outperforms related works in predicting multi-dimensional human preferences across three datasets, which demonstrates the generalization of our method.

Acknowledgements: We sincerely thank Zhuang Li, Lingyu Zou, and Peihan Li for their valuable discussions and feedback.

References

[1] Prompthero. https://prompthero.com/. 4

[2] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021. 3

[3] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 3

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

[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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 1, 3

[6] 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. 1, 2, 3, 5, 6, 7

[7] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *arXiv preprint arXiv:2306.04717*, 2023. 2, 3

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 12888–12900. PMLR, 2022. 2

[9] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 2, 4

[10] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 3

[11] 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 2, 5, 6, 7

[12] 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1, 3, 6

[13] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. *URL https://arxiv.org/abs/2204.06125*, 7, 2022. 3

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 3

[15] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3

[16] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016. 1, 3

[17] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 6, 7

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017. 7

[19] 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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 893–911. Association for Computational Linguistics, 2023. 2, 4

[20] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *CoRR*, abs/2306.09341, 2023. 1, 2, 3, 6, 7

[21] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 2023. 2, 3, 5, 6, 7

[22] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-

to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 1, 2, 3, 5, 6, 7

[23] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022. 2, 3, 4

[24] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. *arXiv preprint arXiv:2303.12618*, 2023. 1, 2, 3

[25] 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. 3