

Article

Not peer-reviewed version

Imaginique Expressions: Tailoring Personalized Short-Text to Image Generation Through Aesthetic Assessment and Human Insights

[Yitian Wan](#)^{*}, Luwei Xiao, [Xingjiao Wu](#), [Jing Yang](#)^{*}, [Liang He](#)

Posted Date: 21 October 2024

doi: 10.20944/preprints202410.1550.v1

Keywords: text-to-image; personalized image aesthetic assessment; human feedback



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Imaginique Expressions: Tailoring Personalized Short-Text to Image Generation Through Aesthetic Assessment and Human Insights

Yitian Wan , Luwei Xiao, Xingjiao Wu, Jing Yang * and Liang He

Department of Computer Science and Technology, East China Normal University, Shanghai 200062, China

* Correspondence: 51255901006@stu.ecnu.edu.cn(Y.W.); jyang@cs.ecnu.edu.cn(J.Y.)

Abstract: Text-to-image tasks have gained significant progress, where vivid images are generated from detailed text descriptions. However, existing studies overlook scenarios where the provided text is sparse, neglect the investigation of human preferences, and fail to acknowledge the diversity of aesthetic opinions. To address those issues, we develop a methodology called personalized short-text-to-image generation through aesthetic assessment and human insights. We develop a Personality Encoder (PE) to extract personal information and establish the Big-Five personality traits-based Image Aesthetic Assessment model (BFIAA) for predicting specific human aesthetic preferences. Utilizing the BFIAA model, we fine-tune the Stable Diffusion model to align it more closely with human preferences. Our experiments demonstrate that our BFIAA model can truly reflect human aesthetic preference and the adapted generation model can generate personalized images more preferred by humans.

Keywords: text-to-image; personalized image aesthetic assessment; human feedback

1. Introduction

Recently, the incredible progress of diffusion models has boosted outstanding performance in text-to-image generation tasks, where those models have achieved distinctive results on automatic evaluation metrics, such as Inception Score (IS) [1] and Fréchet inception distance (FID) [2].

Nevertheless, a contributing factor to the development of state-of-the-art (SOTA) models is the inclusion of elaborate textual descriptions, as exemplified by the passage, “a blonde woman with a rag doll cat sitting next to each other on a bench, cyberpunk art by monet, trending on cgsociety, retrofuturism, reimagined by industrial light and magic, darksynth” [3]. This descriptive narrative provides an extensive account of the target image. It not only determines the image subject and the background but also expresses the image style, wearing details of the subject and auxiliary objects to render to target atmosphere. Nevertheless, when users have no mature ideas or limited-expression ability, short-text to image generation is more pragmatic in real-life applications. In that case, one problem is that we need to generate images from concise textual inputs rather than extensive and detailed textual descriptions.

Simultaneously, as aforementioned automatic metrics focus exclusively on assessing image texture instead of image content [3], they fail to discern human preferences. Consequently, images that receive favorable ratings based on automatic evaluation metrics may not necessarily align with human preference. In that case, to make further study on text-to-image generation tasks, another challenge is that we need to try to make the generated images align well with human aesthetic preference. Existing studies [3,4] adopt the concept of Human-in-the-Loop (HITL) [5], employing a reward model to refine pre-trained text-to-image models to align with human preference. The reward model commonly encompasses a comprehensive large-scale visual-text model, such as CLIP [6]. However, it is doubtful whether CLIP can extract aesthetic features that affect human preference since it only extracts general image features, lacking the explicit identification of aesthetic attributes. As shown in Figure 1, the generated images exhibit obvious discomfort areas that conflict with common sense. For instance, the yellow area in the first image seems to be weird, and it should be the same style as the surrounding areas conventionally.



Figure 1. Generated images within short text easily contain discomfort regions, which conflict with human preference.

Furthermore, there exists a widely recognized consensus that human aesthetics is inherently subjective and diverse. As indicated in prior research [7], individuals with distinct Big-Five personality traits (Openness (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Neuroticism (N)) exhibit varying aesthetic preference for the same image. However, a dearth of studies exists regarding the generation of personalized images from textual inputs. Consequently, it becomes imperative to adaptively model human preference, taking into account the profound semantic nuances inherent in Big-Five personality trait scores.

To mitigate the aforementioned issues, we tailor personalized short-text to image generation via aesthetic assessment and human insights. Following previous work [4], we define short-text as a natural language sentence that consists of the subject word and a combination of specified colors, counts, and background phrases, containing no more than ten characters in total. To learn personalized human aesthetic preference, we take two steps to train a Big-Five personality traits based Image Aesthetic Assessment model (BFIAA). Firstly, we set a five-class text-classification pre-training task to get our novelty proposed Personality Encoder (PE) model, which is made up of a pre-trained language model and afterward layers, to learn the semantic embedding of Big-Five personality traits. Secondly, we train the BFIAA model with personal traits. The BFIAA model takes TAnet [8] as the backbone, adds PE as the personality module, and adopts a cross-attention feature fusion module to fuse aesthetic features with personality embedding to predict the score of the input image. To make generated images align with personalized preference, inspired by previous studies [3,4], we employ the HITL [5,9] framework, but leverage the personal image aesthetic assessment model rather than general large-scale visual-text models [6,10] like CLIP to make pre-trained Stable Diffusion [11] model adapted via LoRA [12].

The proposed PE component achieves an accuracy of 98.1% on the text-classification task. The personalized aesthetic assessment model BFIAA attains state-of-the-art (SOTA) performance on publicly available datasets PARA [7]. Images generated by the text-to-image model are fine-tuned accordingly, yielding results aligning with personalized human preference. Additionally, we conducted necessary ablation experiments and comparative studies, providing evidence of the effectiveness of each proposed strategy. Our contributions are as follows:

- **Use HITL to generate personalized images:** To our best knowledge, this paper is the first to exploit an aesthetic assessment model to finetune the image generation model to get results more in line with human aesthetic preference.
- **Build the Personality Encoder:** We embed the personality traits score distributions into the natural language to get sentences, exploiting them to pre-train a transformer-based model on the text classification task, making the model output the correct personality category. The model will be utilized in the latter training of personalized image aesthetic assessment.

- **Develop Big-Five personality traits based Image Aesthetic Assessment model:** We concurrently input personality traits and image features into the pre-trained PE and the backbone of the aesthetic model, and we introduce the cross-attention fusion module to amalgamate features from different modalities. Subsequently, these features are utilized for mapping, ultimately yielding aesthetic scores.

2. Related Work

2.1. Text-to-Image Generation Models

Text-to-image generation has been a hot topic for a long time. Early on, researchers primarily employed the Generative Adversarial Network (GAN) framework [13], utilizing a generator-discriminator structure for image generation. One of the most representative models was the GAN-CLS [14] introduced in 2016, marking the first GAN model to achieve visually plausible results. With the rise of the Transformer [15] architecture in natural language processing tasks, Transformer-based auto-regressive models [16–18] have emerged in the field of image generation. This approach predicted each image token based on each text token using the Transformer architecture.

In recent research, the diffusion model [11,19,20] architecture has become a trend in image generation models. This method utilized a process based on Markov chains [21], modeling from a Gaussian distribution [22] to progressively denoise images, generating images that conform to textual descriptions from random Gaussian noise [22]. The most representative model in this category is the Stable Diffusion [11] model, leveraging the capabilities of large language models such as BERT [23] to progressively denoise images in latent space, reducing the parameter count while maintaining the authenticity of generated images. The model achieved optimal performance on automated metrics such as FID [2] and IS [1].

Building upon this, generating images that further align with human preference has become a focal point of researchers' attention. Existed studies [3,4,24] often used the thought of Human-in-the-Loop (HITL) [5], and their reward models were often general large-scale visual-text models such as CLIP [6] finetuned on datasets annotated by human choices. Inspired by them, we use the HITL framework, select the Stable Diffusion model as the basic model, but choose an image aesthetic assessment model as the backbone reward model to guide the basic model due to the uncertain human-aesthetics extraction ability of CLIP.

2.2. Image Aesthetic Assessment Task

Image Aesthetic Assessment (IAA) is a critical aspect of image and art analysis, involving the subjective assessment of visual appeal and artistic quality. In recent years, the exploration of computational models for aesthetic judgment has gained substantial attention within the field of computer vision. This task extends beyond traditional image classification, requiring models to discern intricate visual patterns and inherent artistic qualities.

The aesthetic assessment encompasses diverse factors [7], such as composition, color harmony, and overall visual coherence, making it a multidimensional challenge. Compared with general visual-text models that extract the overall features, IAA models specially learn image aesthetics from various image attributes, involving human opinions as well. [25] learned aesthetics from image composition and image styles to figure out the reasons why images attracted human. [26] directly modeled human opinions toward images to explicitly obtain human aesthetic preference. [8] believed that images with different themes had different evaluation criteria, so it studied image aesthetics with corresponding criteria to analyze human assessments.

In that case, we can see that the IAA task is subjective which can reveal human aesthetic preference towards images, and IAA models are directed against the aesthetic attributes to learn human aesthetics. As a subjective task revealing human aesthetic preferences, IAA models serve as reward functions in Human-in-the-Loop (HITL) frameworks for fine-tuning image generation models.

2.3. Connections between Big-Five Personality Traits and Human Aesthetic Preference

The Big-Five personality traits [27], also known as the Five-Factor Model (FFM), is a widely accepted and influential framework for understanding and describing human personality. This model categorizes personality traits into five broad dimensions, which are Openness (O), Conscientiousness (C), Extroversion (E), Agreeableness(A), and Neuroticism(N). O reflects the extent to which an individual is open-minded, imaginative, curious, and willing to engage in novel experiences. C measures the degree of organization, responsibility, dependability, and goal-oriented behavior in individuals. E represents the degree to which a person is outgoing, sociable, energetic, and enjoys interacting with others. A gauges one's tendency to be cooperative, empathetic, kind, and considerate of others. N measures emotional stability and resilience to stress.

[7] showed that people with different personality traits have different concerns about aesthetic attributes and give diverse scores to the same image. For example, people with higher N scores tend to focus more on composition, while people with higher E scores tend to pay more attention to image content. Therefore, it's of great necessity to learn personalized aesthetic preference, and obviously, the traditional visual-text models can not solve the problem, we need to make great use of the personality traits scores and train a personalized IAA model instead.

3. Methods

The overall architecture of our framework is illustrated in Figure2. It consists of four parts: the pre-trained Stable Diffusion model, the Personality Encoder (PE), the proposed BFIAA model, human insights respectively. Drawing inspiration from the Human-in-the-Loop (HITL) concept, we employ the Stable Diffusion as the foundational model. We then integrate the BFIAA model, which incorporates human personalized information encoded by PE as a reward function, thereby enabling the foundational model to adapt accordingly, so that we can get the final personalized image generation model. The subsequent sections will provide a detailed exposition of each process within our framework.

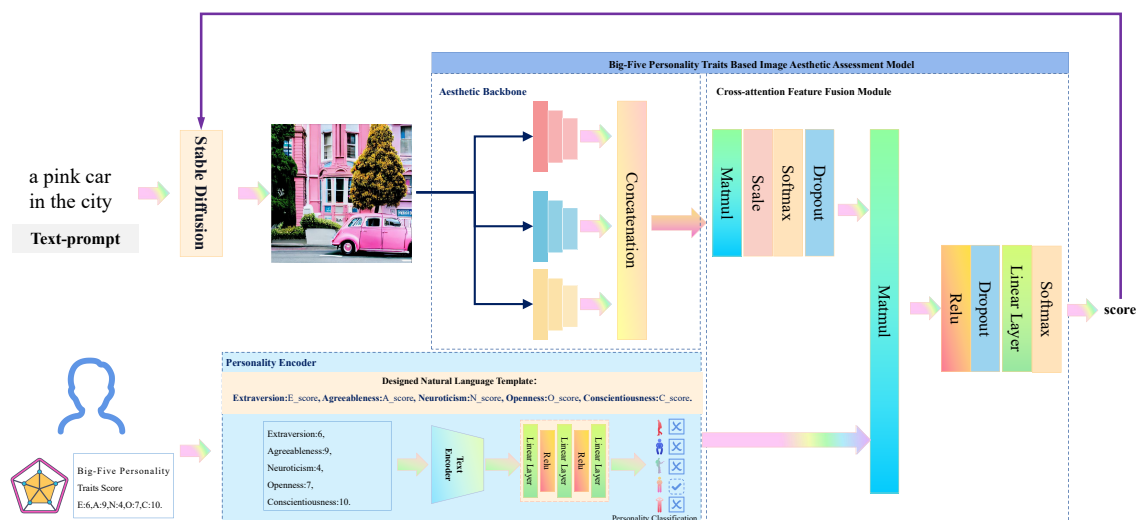


Figure 2. Overall framework of our proposed method.

3.1. Personality Encoder

3.1.1. Overall Architecture

Firstly, we design a natural language template, as shown in Figure2. We integrate the score distributions of the five dimensions (O, C, E, A, N) into natural language templates, transforming them into personality text $text_p$. Subsequently, our objective is to construct a model for a text-classification

pre-training task, which compels the model to proficiently extract semantic features. The architecture of the model is illustrated in the blue area of Figure 2, comprising a pre-trained transformer-based language model as the pre-trained text encoder, denoted as all-mpnet-base-v2, and a Multilayer Perceptron (MLP) module as the adapter layers. The whole process can be formulated as:

$$\begin{aligned} s_{PE} &= \text{Adapter}(\text{TextEncoder}(\text{text}_p)) \\ p_c &= \text{Linear}(s_{PE}) \end{aligned} \quad (1)$$

where s_{PE} refers to personality embedding. p_c is the predicted class label. Notably, the final Personality Encoder is derived from the aforementioned model by excluding the last linear layer of the adapter layers designed for classification, and the s_{PE} will then present the embedding of the deep meanings in personality traits used in personalized image aesthetic assessment step described in the next section.

3.1.2. Data Preparation

Data for the pre-training text classification task is combined with pairs made up of a personality text and a corresponding label, which is prepared as follows: First, we extract personal information from the PARA [7] dataset. Then, we set every user a ground-truth personality traits label depending on the preference of the score distribution on the Big-Five personality traits. Last, we embed the score distribution into the template mentioned above to get the corresponding personality text, and we take the dimension with the highest score as the label.

3.1.3. Experiment of the Pretraining Task

We choose the Adam optimizer and cross-entropy loss to retrain our model with the prepared dataset, the loss function can be formulated as:

$$\text{loss}_{PE} = - \sum_i \text{label}_i \cdot \log(p_i) \quad (2)$$

where p_i represents the predicted label and label_i represents the ground-truth personality label. As a result, we finally obtain a classification accuracy of 98.1%. And we do ablation study to validate the effectiveness of our chosen pre-trained model and learning rate of PE, it will be shown in the Experiments section.

3.2. Big-Five Personality Traits Based Image Aesthetic Assessment Model

3.2.1. Overall Architecture

As shown in Figure 2, the BFIAA model comprises the aesthetic backbone, the personality encoder, and the cross-feature fusion module. Regarding the aesthetic backbone, since one of the key factors for human evaluation is the thematic information of an image, we opt for TANet [8] to derive the overall aesthetic features f_{aes} . This selection is driven by TANet's achievement of SOTA performance in extracting significant image theme features, and then we can achieve comprehensive understanding of the given image through three aspects: thematic feature branch, image aesthetic feature branch, and color feature branch. Among those branches, image aesthetic feature branch is achieved by a pre-trained MobileNet and afterwards layers, and color feature branch is composed of a RGB attention module.

For the personality encoder, taking personality text as input, we use our pre-trained PE introduced in Sec3.1 to obtain s_{PE} , which refers to the embedding of deep meanings in personality traits. The process of the extractions can be formulated as:

$$s_{PE} = PE(\text{text}_p). \quad (3)$$

Following this, the cross-feature fusion module integrates the personality embedding s_{PE} with image aesthetic features f_{aes} from aesthetic backbone, resulting in the derivation of personalized aesthetic features f_p for image assessment. Inspired by previous works that fuse information from different modules, as shown in figure 2, we adapt the modified cross attention mechanism as the basic architecture. To enhance the personalization of these features, we implement a dropout operation subsequent to the softmax function. This addition aims to diminish co-adaptation between attention weights and values, thereby refining the feature's individualized characteristics. Furthermore, in order to better map the personality enhanced image feature into aesthetic score y_{score} , we apply an adapter module consists of relu function, dropout operation, linear layer and softmax function. The relu function is used to introduce nonlinear feature to handle complex features, and the linear layer is used to do regression tasks like score generation. The whole process can be formulated below:

$$f_p = dropout(softmax(\frac{KQ^T}{\sqrt{d_k}}))V \quad (4)$$

$$y_{score} = softmax(Linear(dropout(ReLU(f_p)))) \quad (5)$$

where Q and K present the overall aesthetic features and personality embeddings. V equals to K , and d_k refers to dimension value of input features. Ultimately, the final score y_{score} is obtained as reward score to adjust the stable diffusion model.

3.2.2. Data Preparation

We transform the PARA [7] dataset into a Big-Five personality traits oriented image aesthetic assessment dataset, which contains 5 sub-datasets where participants label images in the same sub-dataset with the same personality traits. We make preparations as the following steps: Firstly, we find every image and its corresponding annotations and divide all the annotations into 5 groups depending on the annotators' preference for personality traits. Secondly, we construct each personality text by embedding the average score distribution into the template. Lastly, we build 5 sub-datasets that stand for O, C, E, A, and N personal traits, and each sub-dataset contains pairs that involve image, corresponding average aesthetic score, and the personality text.

3.2.3. Experiment on Prepared Data

We choose the Adam optimizer and Mean Squared Error (MSE) loss to train our BFIAA model:

$$loss_{BFIAA} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

where y_i means the ground-truth score and $\hat{y}_i$ is the predicted score. Eventually, we achieve SOTA performance on all five sub-datasets. The detailed results and ablation study will be discussed in the Experiment section.

3.3. Personalized Image Generation Model

3.3.1. General Architecture

Human-in-the-Loop refers to a strategy that integrates human feedback into machine learning and artificial intelligence systems. In this approach, the reward function guides the base model to align with human preferences by learning from them. Inspired by this idea, we applies a personalized aesthetic evaluation model as a specific reward function in the task of short-text-to-image generation, mapping human aesthetic preferences to quantify the alignment between generated images and human aesthetic tastes. By fine-tuning the base model through LoRA on a specific training dataset, the base

Stable Diffusion model iteratively improves its image generation capabilities according to human preferences, resulting in a personalized short-text-to-image generation model. The following sections will detail the construction process of the specific dataset and the steps for LoRA fine-tuning to get the final personalized image generation model.

3.3.2. Data Preparation

The training dataset consists of over 27,000 entries, with each entry comprising three components: a short text prompt, multiple generated image-aesthetic score pairs. For the construction of short text prompts, inspired by previous research [8], we designed three categories of prompts: quantity, color, and background. Each category includes common elements corresponding to that category, and we also created a set of terms that represent common objects with varying visual characteristics. Within each category, we generated prompts by combining one word or phrase from the category with an object. For example, combining "white" with "bird." Additionally, we explored combinations of two or three categories, such as "two birds in a city."

For each short text prompt, we generated three images using the base Stable Diffusion model and assigned human aesthetic scores to these images through our personalized aesthetic evaluation model. To differentiate between various personality traits, we utilized the BFIAA model to assign distinct aesthetic scores to each image.

3.3.3. Adapt Stable Diffusion with BFIAA Model via LoRA

To reduce training costs while effectively aligning the base model with human preferences, we use the LoRA[12] approach to fine-tune the base Stable Diffusion model. When fine-tuning with LoRA, the parameters of the original model remain frozen, meaning that we do not significantly alter the model's fundamental structure, ensuring its basic capabilities remain intact. Meanwhile, the parameter matrices corresponding to the key, query, and value are enhanced by introducing low-rank residuals:

$$\begin{aligned} \mathbf{W}'_K &= \mathbf{W}_K + \Delta \mathbf{K} \\ \mathbf{W}'_Q &= \mathbf{W}_Q + \Delta \mathbf{Q} \\ \mathbf{W}'_V &= \mathbf{W}_V + \Delta \mathbf{V} \end{aligned} \quad (7)$$

where $\mathbf{W}'_K, \mathbf{W}'_Q$ and $\mathbf{W}'_V$ represent the new parameter matrices for the key, query, and value, respectively, and $\Delta \mathbf{W}'_K, \Delta \mathbf{W}'_Q$ and $\Delta \mathbf{W}'_V$ represent the corresponding low-rank residual matrices. To obtain the low-rank residual matrices, we first initialize them as low-rank matrices. Then, we define a loss function to evaluate the difference between the model's predictions and the target outputs. During training, the gradient of the loss function with respect to the low-rank matrices is calculated via back propagation, and an optimizer is used to update them:

$$\Delta \theta \leftarrow \Delta \theta - \eta \cdot \nabla_{\Delta \theta} L \quad (8)$$

where θ is the learning rate and L is the loss function. At this point, LoRA captures the adaptability of the model to specific tasks through the low-rank matrices, thereby enhancing its capacity. Through this mechanism, the obtained low-rank residual matrices can be seamlessly integrated into the base model after training, ensuring the model maintains its original stability and efficiency. Additionally, to reduce undesirable regions in the generated images and incorporate personalized descriptions that align more closely with individual aesthetic preferences, we define the loss function L as follows:

$$L = loss_{MSE} \cdot \sigma(score_{reward}) \quad (9)$$

$$loss_{MSE} = \frac{1}{n} \sum_{i=1}^n \|pred_{noise_i} - target_{noise_i}\|^2 \quad (10)$$

$$score_{reward} = BFIAA(x_i) \tag{11}$$

where $BFIAA$ represents the personalized aesthetic evaluation model based on the Big Five personality traits, and $score_{reward}$ is the reward score corresponding to the real image x_i , i.e., the personalized aesthetic score. σ denotes the sigmoid function, used to scale the score to the range of 0 to 1, while $pred_{noise_i}$ and $target_{noise_i}$ represent the predicted and true values of the noise to be added to the image at time step i in the diffusion model. This loss function shows that when an image does not meet human preferences, even if the predicted noise and true noise differ significantly, the final loss value will not be excessively large due to the control of the reward score. Conversely, when an image aligns with human preferences, even if the difference between the predicted and true noise is not very large, the loss function will still yield a larger value due to the scaling of the reward function. Thus, through continuous iteration, the model gradually generates images that increasingly align with human preferences.

4. Experiments and Analysis

4.1. Results of the Personality Encoder

We utilized a text-classification pre-training task to enable a deep learning model to learn the meanings of personality traits by analyzing corresponding score distributions, then we take prediction accuracy as the evaluation metric, and finally achieve 98.1% accuracy, which indicates that our personality encoder can obtain embedding that contains adequate information to categorize the personality text into an accurate class.

During our research, we conducted ablation studies on the proposed Personality Encoder (PE). As shown in Table 1, We experimented with different pre-trained models, afterwards layer structures, learning rates, and embedding dimensions to determine the best combination for achieving optimal results. As for pre-trained models, we try a classical text encoder used for multimodal tasks like CLIP and compare it with the selected all-mpnet-base-v2 model that achieves the best results on several datasets used for pure natural language processing tasks. As we can see in the second row and the third row, the selected all-mpnet-base-v2 achieves better results, and the accuracy is improved by 12.5%. In terms of afterward layer structure, we use MLP in our PE, and to validate its effectiveness, we use a convolutional structure to replace the original part to see which one is better. As we can see in the first two rows, the MLP structure performs a lot better. Then, we try different combinations of the learning rate and embedding dimensions, and setting the learning rate to 0.001 and the embedding dimension to 30, the final result comes to the highest point of 98.1%.

Table 1. Experiment results of the text-classification pretraining task on Personality Encoder, where PM refers to the pre-trained text encoder, AL refers to the adapter layers after the pre-trained model, ED refers to the embedding dimension value and LR refers to the learning rate when training.

PM	AL	ED	LR	Accuracy
CLIP_ViT_H_14	ConvNet	100	0.0005	54.3%
CLIP_ViT_H_14	MLP	100	0.0005	76.8%
all-mpnet-base-v2	MLP	100	0.0005	86.4%
all-mpnet-base-v2	MLP	100	0.001	94.6%
all-mpnet-base-v2	MLP	30	0.001	98.1%

4.2. Results of Our BFIAA Model

Following previous works[7,8,25], we select SRCC, PLCC, and Accuracy as the evaluation metrics. SRCC stands for the Spearman’s Rank Correlation Coefficient (SRCC), which measures the monotonic relationship between the ground-truth and the predicted scores, ranging is [-1, 1]. And PLCC refers to Pearson’s Linear Correlation Coefficient (PLCC), which measures the linear correlation between the

ground-truth and the predicted scores, also ranging is [-1, 1]. The quantitative results are shown in Table 2, where TANet gets 0.711 on PLCC and 0.652 on SRCC and BFIAA gets 0.761 on PLCC and 0.718 on SRCC. The results of baseline of PARA are gained from [7], which are 0.751 and 0.705 respectively.

Table 2. Quantitative PIAA task results.

Evaluation Metrics	Baseline in PARA[7]	TANet	BFIAA (Ours)
SRCC	0.705	0.752	0.718
PLCC	0.751	0.711	0.761

Additionally, as shown in Table.3, we make ablation studies to demonstrate the value of the Personality Encoder module and cross-attention fusion module. Comparing the first part with the second part, we can observe that the inclusion of PE leads to an improvement in the results of all five sub-datasets across all metrics. Additionally, when comparing the second and third parts, we can see a noticeable improvement in the performance of all sub-datasets in terms of SRCC and PLCC, except for accuracy, which indicates that the cross-attention fusion module is of good value.

Table 3. Abaltion study results of the BFIAA model, where PE means the proposed Personality Encoder module and CA means the cross-attention fusion module.

Model	Personality Traits	SRCC	PLCC	Accuracy
w/o PE, CA	O	0.752	0.801	98.9%
	C	0.784	0.842	98.4%
	E	0.516	0.582	98.0%
	A	0.693	0.761	99.4%
	N	0.516	0.571	98.6%
w/o CA	O	0.756	0.815	99.0%
	C	0.803	0.862	98.9%
	E	0.522	0.585	98.5%
	A	0.743	0.787	99.4%
	N	0.598	0.629	98.2%
BFIAA	O	0.780	0.821	99.0%
	C	0.816	0.862	98.2%
	E	0.628	0.670	98.2%
	A	0.755	0.796	99.3%
	N	0.613	0.656	98.4%

4.3. Results of the Personalized Image Generation Model

We divided our experiments into two parts, the quantitative experiments and the qualitative experiments. Quantitative results are shown in Table 4, we take the CLIP score to validate the alignment between text and image and MSE loss between latent noises in the denoising process as evaluation metrics. In Table 4, our model achieves better performance than the baseline model on both metrics, improving by 0.13% and 28.8% respectively.

Table 4. Quantitative results on automatic evaluation metrics, where O, C, E, A, N refer to results of different personality traits.

Model	CLIP score	MSE loss
Baseline	0.383	0.132
O	0.383	0.0868
C	0.385	0.0868
E	0.383	0.0868
A	0.383	0.0868
N	0.384	0.0868

As for qualitative experiment, We have divided it into two parts. Firstly, we aim to validate our model’s ability to produce images with fewer anomalies in response to short-text inputs. To achieve this, we will compare the images generated by our model with those generated by the baseline introduced in [4]. Secondly, we seek to establish that our model can generate a range of personalized images tailored to individual preferences. We will do this by comparing the image outputs that align with diverse aesthetic preferences.

As shown in Figure 3, we discover that our model generates images with better quality and aesthetics than the baseline model. Our model outperforms the baseline model in several ways. In the first three columns, our images have no white boxes indicating discomfort areas, and they appear more vibrant. In the next two columns, our images have more detailed and colorful backgrounds, while the subjects remain of high quality. Lastly, our model successfully generates the required subject, "purple cake," and the background scene looks more artistic and layered.

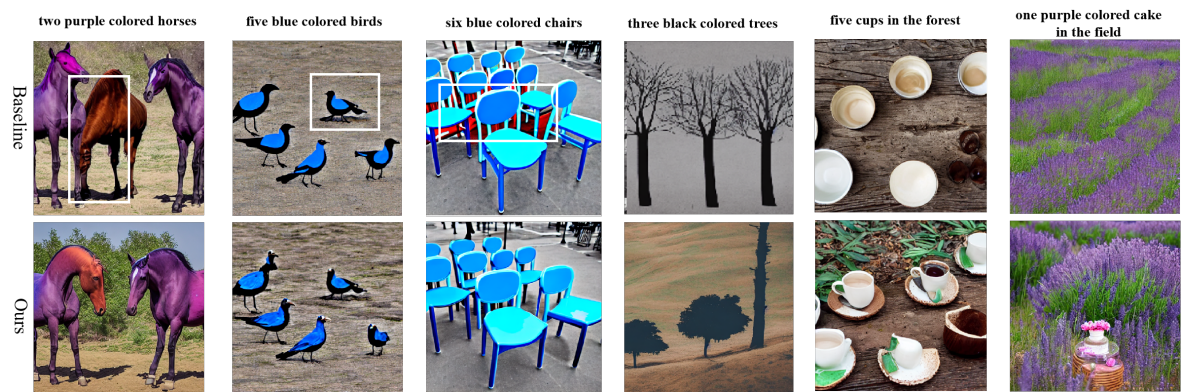


Figure 3. Comparison between our generated images and images generated from baseline model in [4].

Figure 4 presents images generated by our model for individuals with openness and conscientiousness personality traits. Overall, each image corresponds to the keywords or main subjects of the short text prompts, with no regions that contradict commonsense. Additionally, there are notable differences in the images preferred by individuals with openness versus conscientiousness.



Figure 4. Generated images comparison between personality trait O aesthetically preferred and personality trait C aesthetically preferred.

From the perspective of color and details, the images in the first row feature richer and more varied colors, such as the sheen of the car and the intricate markings of the clock. These detailed elements provide more visual exploration space, catering to the curiosity of individuals with high openness. In contrast, the second row of images displays softer tones, with lower color saturation and weaker contrast between objects. For example, the background details of the vase are simpler and quieter, with minimal visual stimulation. This muted and orderly tone aligns with the preference for structure and clarity typical of conscientious individuals.

In terms of composition, the scenes in the first row exhibit stronger contrasts, such as the car and the building, especially the body of the clock standing out prominently and the cake decorations appearing more unusual. This sense of contrast appeals more to individuals with high openness, who prefer novel and unconventional elements. Meanwhile, the composition in the second row is more balanced, with objects arranged in an orderly manner and no striking or unusual items, reducing visual complexity. This structured and precise arrangement corresponds to the preference of conscientious individuals for orderliness and clarity.

Figure 5 presents images generated by our model for individuals with extraversion and agreeableness personality traits. From a general aesthetic perspective, each image is harmonious and free of regions that violate common sense or cause discomfort. In terms of personalized preferences, the images corresponding to different personality traits exhibit variations in details, colors, and backgrounds.



Figure 5. Generated images comparison between personality trait O aesthetically preferred and personality trait C aesthetically preferred.

From the perspective of visual complexity, the images in the first row contain multiple focal points, such as the arrangement of vehicles, the layered decorations on the cake, and the diagonally extending branches of the plant. This creates a rich and diverse visual experience, which strongly appeals to individuals with high extraversion. In contrast, the images in the second row have lower visual complexity. The scenes appear more orderly and avoid excessive visual noise, such as the more neatly arranged vehicles and the more neatly pruned branches. Individuals with high agreeableness tend to prefer such simpler and more harmonious images, which do not induce excessive emotional stimulation.

Figure 6 presents images generated by our model for individuals with conscientiousness and neuroticism personality traits. It can be observed that the images in the second row exhibit softer colors with lower contrast. This low-stimulation color palette helps reduce tension and creates a more

relaxed feeling, particularly through the more natural and comfortable postures of the animals, which aligns with the need of individuals high in neuroticism for gentle, calming visual experiences.



Figure 6. Generated images comparison between personality trait C aesthetically preferred and personality trait N aesthetically preferred.

Furthermore, the images in the second row convey a softer overall style. The colors of the cake are slightly muted, and the postures of the horse and dog are more relaxed. Additionally, the background elements, such as the green trees and outdoor setting, evoke a serene, natural atmosphere. This harmonious scene helps to alleviate the heightened sensitivity to external stimuli often experienced by individuals with high neuroticism.

As shown in Figure 7, the overall performance of our model demonstrates relatively clear differences among the five columns of images representing personalized preferences. This indicates that our model can detect aesthetic differences between various personality traits. In the first row, the base of the clocks shows different hollow carving designs, with the neuroticism clock having a flat bottom, while only the conscientiousness clock features a square outline.

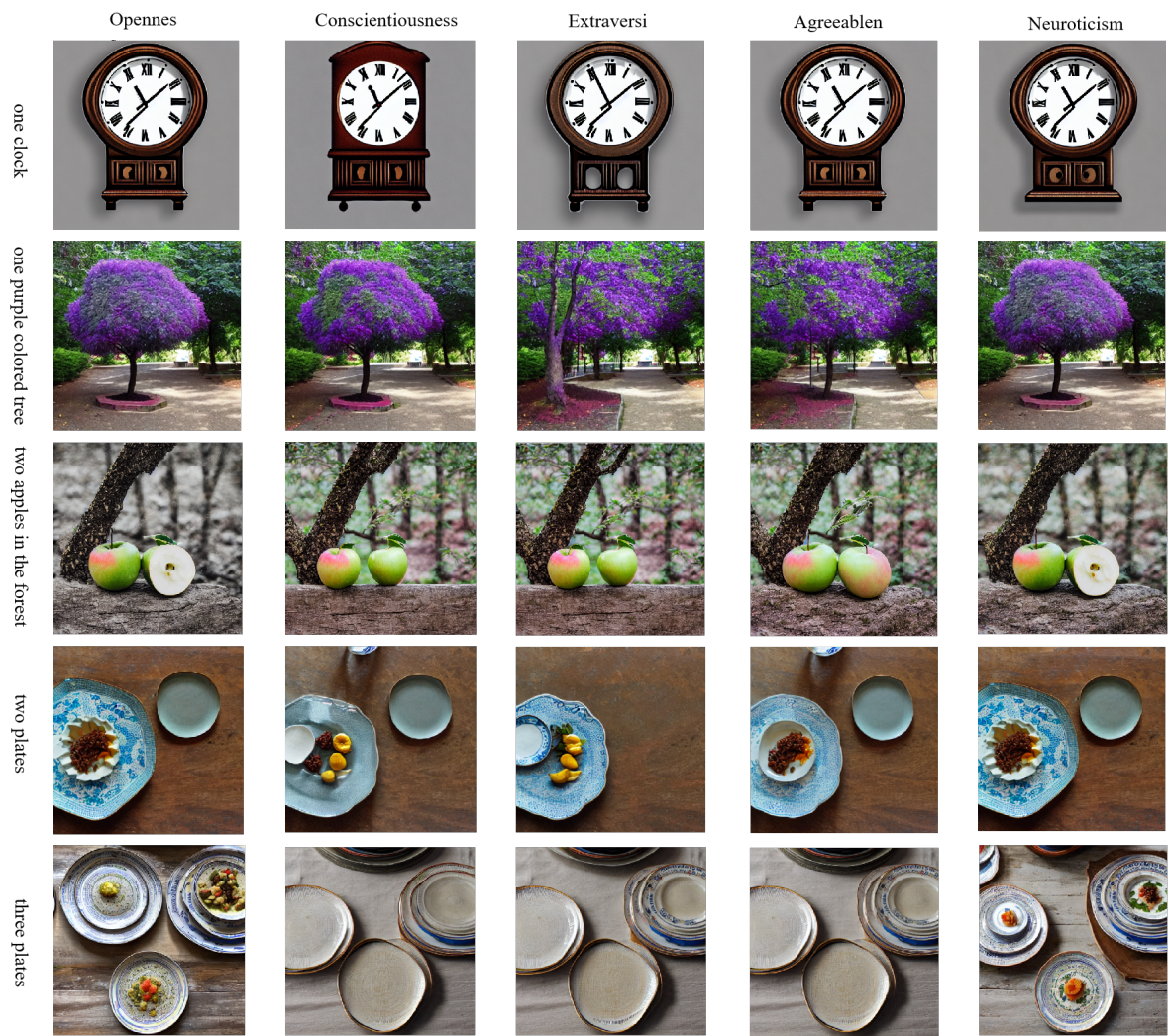


Figure 7. Generated personalized images comparison between the five personality traits.

In the second row, the bases of the trees in the extraversion and agreeableness images are relatively larger squares compared to those in other images, and the conscientiousness image contains an additional tree branch compared to the agreeableness image. In the third row, only the openness and neuroticism images include half of an apple; notably, only the openness image has a distinct black-and-white background, whereas the other images feature colored backgrounds.

In the final row, it is evident that each plate contains different types of food, with some plates featuring additional yellow fragments, while others have brighter blue patterns. We observed that while there is a certain degree of similarity between images corresponding to different personality traits, subtle differences in detail are still noticeable. Furthermore, when the text length remains constant, the larger the numbers included in the text, the greater the differences between the generated images.

From this, it can be seen that our adapted personalized short-text image generation model not only achieves the effect of reducing abnormal areas in terms of common sense aesthetics but also distinguishes itself in terms of personalized aesthetics while maintaining overall consistency between the text and image.

5. Future Discussion

This paper has established a new angle to obtain image generation model from short texts that align well with personalized aesthetic assessments. Though we propose a personality encoder to

learn deep semantic meanings of the big five personality traits, it has some aspects to get improved in the future. For example, we can add commonsense knowledge from knowledge database or large language models to enrich the characteristic information. Furthermore, contrastive learning can be used to learn specific personal-traits preferred aesthetic features. Last but not least, we can use an adapter or projection networks after the image generation model and freeze the image generation model instead of LoRA to train with lower costs.

6. Conclusions

This paper focuses on the challenges associated with generating images from short texts, including issues with discomfort image areas and the neglect of individual aesthetic preferences. To address these challenges, we tailor the personalized short text to image generation through aesthetic assessment and human feedback, drawing inspiration from the HITL framework. We begin by developing a Personality Encoder, which uses a pre-training task to extract profound semantic features related to personality based on Big-Five score distributions. We then introduce the PIAA model, or the Big-Five Personality Traits-based Image Aesthetic Assessment model, to better understand human aesthetic preferences in the context of their personalities. Finally, we refine the Stable Diffusion model with guidance from the PIAA model, resulting in an image generation model that is better aligned with personalized aesthetic preferences. Our hope is that this approach will allow users to generate high-quality images that accurately reflect their individual tastes with minimal effort.

Author Contributions: Conceptualization and methodology, Y.W.; software and validation, Y.W.; formal analysis and investigation, Y.W.; resources, Y.W.; data curation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, L.X. and X.W.; visualization, L.X. and X.W.; supervision, J.Y. and L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Acknowledgments: Thanks to the editor and anonymous reviewers for their suggestions and comments that ameliorate our work.

Conflicts of Interest: The authors declare no conflicts of interest..

References

1. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. *Advances in Neural Information Processing Systems (NeurIPS)* **2016**, *29*.
2. 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. *Advances in Neural Information Processing Systems (NeurIPS)* **2017**, *30*.
3. Wu, X.; Sun, K.; Zhu, F.; Zhao, R.; Li, H. Better aligning text-to-image models with human preference. *International Conference on Computer Vision (ICCV)*, 2023.
4. Lee, K.; Liu, H.; Ryu, M.; Watkins, O.; Du, Y.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Gu, S.S. Aligning text-to-image models using human feedback. *arXiv* **2023**.
5. Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; He, L. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* **2022**, *135*, 364–381.
6. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others. Learning transferable visual models from natural language supervision. *Proceedings of International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.
7. Yang, Y.; Xu, L.; Li, L.; Qie, N.; Li, Y.; Zhang, P.; Guo, Y. Personalized image aesthetics assessment with rich attributes. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19861–19869.
8. He, S.; Zhang, Y.; Xie, R.; Jiang, D.; Ming, A. Rethinking image aesthetics assessment: Models, datasets and benchmarks. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 942–948.
9. Qiao, N.; Sun, Y.; Liu, C.; Xia, L.; Luo, J.; Zhang, K.; Kuo, C.H. Human-in-the-loop video semantic segmentation auto-annotation. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 5881–5891.

10. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 12888–12900.
11. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.
12. Hu, E.J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; others. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR)*, 2021.
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)* **2014**, 27.
14. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative Adversarial Text to Image Synthesis. *Proceedings of International Conference on Machine Learning (ICML)*, 2016, Vol. 48, pp. 1060–1069.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)* **2017**, 30.
16. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. *Proceedings of International Conference on Machine Learning (ICML)*, 2021, pp. 8821–8831.
17. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**.
18. Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.H.; Murphy, K.; Freeman, W.T.; Rubinstein, M.; others. Muse: Text-to-image generation via masked generative transformers. *Proceedings of International Conference on Machine Learning (ICML)*, 2023.
19. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)* **2020**, 33, 6840–6851.
20. 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. *Proceedings of International Conference on Machine Learning (ICML)*, 2022.
21. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)* **2021**, 34, 8780–8794.
22. Do, C.B. The multivariate Gaussian distribution. *Section Notes, Lecture on Machine Learning, CS* **2008**, 229.
23. Kenton, J.D.M.W.C.; Toutanova, L.K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
24. 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* **2023**.
25. Celona, L.; Leonardi, M.; Napoletano, P.; Rozza, A. Composition and style attributes guided image aesthetic assessment. *IEEE Transaction on Image Processing* **2022**, 31, 5009–5024.
26. Niu, Y.; Chen, S.; Song, B.; Chen, Z.; Liu, W. Comment-guided semantics-aware image aesthetics assessment. *IEEE Transaction on Circuits and Systems for Video Technology* **2022**, 33, 1487–1492.
27. Goldberg, L.R. An alternative “description of personality”: The Big-Five factor structure. In *Personality and Personality Disorders*; 2013; pp. 34–47.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.