Generation of Clothing Patterns Based on Impressions Using Stable Diffusion

J. N. Htoi Sann Ja, Kaede Shiohara, Toshihiko Yamasaki, Miyuki Toga, Kensuke Tobitani, and Noriko Nagata

Abstract Personalized products based on individual preferences have been considered to improve personal well-being and consumer satisfaction. This approach helps reduce waste and conserve resources. With artificial intelligence enabling personalization, consumers can easily access products that match their preferences without the need for specialized knowledge or professional expertise. Advances in artificial intelligence, text-to-image models in particular, have enabled the generation of impressive images from textual descriptions. However, existing models lack the ability to generate images based on visual impressions. In this paper, we propose a text-to-image diffusion model that incorporates visual impressions into the image generation process. Our model extends the stable diffusion architecture by introducing a multi-modal input system that processes text descriptions, pattern images, and quantified visual impressions. Experimental validation confirmed the positive correlation between generated and original images across multiple impression metrics, demonstrating the model's effectiveness in preserving impression-based characteristics. These results suggest that our approach successfully bridges the gap between textual descriptions and visual impressions in image generation.

J. N. H. S. Ja (✉) · M. Toga · N. Nagata
Kwansei Gakuin University, Sanda, Hyogo, Japan
e-mail: sannjajn28@kwansei.ac.jp

M. Toga
e-mail: toga.m@kwansei.ac.jp

N. Nagata
e-mail: nagata@kwansei.ac.jp

K. Shiohara · T. Yamasaki
The University of Tokyo, Bunkyo, Tokyo, Japan
e-mail: shiohara@cvm.t.u-tokyo.ac.jp

T. Yamasaki
e-mail: yamasaki@cvm.t.u-tokyo.ac.jp

K. Tobitani
Institute of Advanced Media Arts and Sciences, Ogaki-shi, Japan
e-mail: tobitani@iamas.ac.jp

X.-S. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1416,
https://doi.org/10.1007/978-981-96-6441-2_14

Keywords Aesthetic · Visual impression · Kansei · Stable diffusion · Image generation

1 Introduction

Personalization in product design plays an important role in modern industries, driven by the growing demand for products that reflect individual preferences and lifestyles. Product designs that reflect user-specific aesthetics improve user satisfaction, enhance emotional connections to products, and contribute to social and economic well-being. Integrating personal aesthetic values into product design can improve user experience and promote sustainable consumption by reducing waste from unwanted products. The fashion industry, in particular, heavily relies on aesthetic appeal, where patterns, textures, and colors influence consumer preferences and emotional responses. Personalizing these aspects through artificial intelligence would enable consumers to easily obtain products that match their tastes without requiring specialized knowledge, making it highly practical.

The rapid development of generative models, particularly text-to-image (T2I) technologies such as Stable Diffusion [1], DALL-E2 [2], and Imagen [3], has enabled the generation of impressive images aligned with textual descriptions. However, while generative technologies allow for customization and creativity, current methods often lack the ability to embed nuanced, subjective human aesthetics, or impressions into the design process. This gap limits their applicability in creating products that reflect deeper emotional and sensory qualities.

To address this challenge, studies on quantifying subjective impressions, developing automatic prediction models for impressions and texture synthesis have been done [4, 5]. However, the generation of images based on impressions as controllable parameters is not established.

In this paper, we propose a pattern image generation model in which the levels of impression values such as “cute”, “bright”, and “cool-looking” evoked from clothing patterns can be simultaneously determined using a stable diffusion model. This technique helps to generate an image that aligns with both the aesthetic values of the users and the textual condition.

2 Related Work

Text-to-image (T2I) diffusion models [1–3] generate images by progressively denoising in an image space or a latent space. These models take text input, encoded using a language or vision language model, such as CLIP, to generate corresponding images. However, relying only on text for conditioning is insufficient in personalized or complex scenarios. For example, general T2I models lack the capability to generate images of unseen individuals.

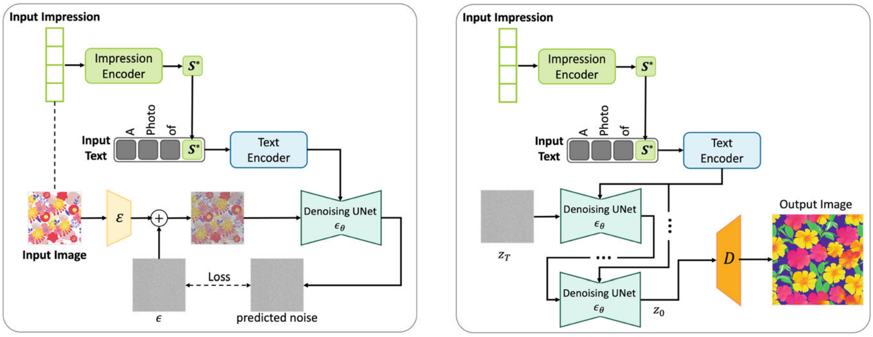
To address this limitation, different types of approach have been developed to personalize the T2I models [6, 7]. For example, the E4T model introduces a two-stage personalization approach, consisting of pre-training an embedded word and adjusting the weight offsets of attention layers on domain-specific datasets followed by the fine-tuning process [6]. This enables the model to adopt novel concepts and generate personalized images from small data samples in a short time.

On the other hand, studies have been conducted on texture [8, 9] and transformation of physical parameters into impressions (vice versa) [4, 5]. Sunda et al. succeeded in constructing an automatic impression estimation model by using the style content of patterned images of clothing as physical parameters. In addition, they developed a pattern search system that aligns closely with human intuition [4]. However, using the same style features and framework for generating new patterned images has not yielded results effective enough to meet expectations.

In this study, our objective is to generate new patterned images that reflect desired impressions by using a personalized T2I model.

3 Proposed Method

Our aim is to encode input impressions as an embedding (S^*) within CLIP's text space, making it possible to generate images with desired impressions guided by text prompts in Stable Diffusion. The overview of the proposed method is visualized in Fig. 1. Our method is based on the Stable Diffusion model (Sect. 3.1) and consists of an impression encoder (Sect. 3.2), where a vector of input impressions is transformed into impression embeddings (S^*), making it unique compared to other models.



(a) Training **(b)** Inference

Fig. 1 Overview of our proposed method

3.1 T2I Model

We use Stable Diffusion (SD) [1] as our base T2I model. SD consists of two encoders, a text encoder and a VAE encoder. The VAE encoder ε compresses images (x) into a latent space, preserving perceptual details while reducing dimensionality. The input text prompt (p) is encoded into embeddings ($\tau(p)$), using the CLIP text encoder (τ). A UNet (ϵ_θ) progressively denoises a noisy latent feature (z_t) at a time step t, guided by a text feature ($\tau(p)$) to predict the added noise ϵ. The model is trained to minimize the difference between actual noise (ϵ) and predicted noise (ϵ), using a mean squared error loss (L_{ldm}) as shown in Eq. 1. Through this training process, SD learns robust semantic relationships between text and images, enabling the generation of high-quality images aligned with textual prompts.

$$L_{ldm} = ||(\epsilon - \epsilon_\theta(z_t, t, \tau(p))||_2^2. \tag{1}$$

3.2 Impression Encoder

We extract the impression feature as a word embedding S^* using the impression encoder. The impression encoder (f_{imp}) is a neural network that processes a vector of impression values (v_x), quantified impressions based on subjective evaluation, corresponding to the input image (x), and converts them into 768-dimensional vectors (Eq. 2).

$$S^* = f_{imp}(v_x). \tag{2}$$

This transformation maps the input to a higher-dimensional space that aligns with the CLIP model's embedding size. Each layer of the network utilizes the LeakyReLU activation function, effectively extracting meaningful features from the impression vectors, enabling seamless integration with the CLIP model. The embedding S^* is then used as a tokenized word to generate impression-enhanced images, as in Fig. 1b.

3.3 Training Process

During training, we input impression values, image, and text prompts. The impression encoder encodes a vector of impression values and the impression feature is extracted as a word embedding S^*, which can then be processed in the text space. The combination of S^* and the input text is encoded by the text encoder. The input image is also encoded by the VAE encoder ε and then a Gaussian noise ϵ is added. All inputs are then processed by the UNet(ϵ_θ) to predict the added noise ϵ (Fig. 1a).

3.4 Inference

During inference, we input impression values and text prompts. As in the training process, the impression feature S^* is incorporated into an input text. Then, denoising is performed to generate an image conditioned on the input impression and text (Fig. 1b).

4 Generation of Images

4.1 Dataset

We conducted experiments using two datasets: (1) floral patterns and (2) general patterns.

Floral Patterns. For the input images, 3,098 floral pattern images with the size of 512 × 512 were prepared (Fig. 2a). For the input impressions, we used 10 evaluation words (Table 2) describing the impressions evoked by the floral patterns. These words were selected through the free description and goodness-of-fit evaluations conducted in previous research [4].

The impressions of the patterns were quantified based on subjective evaluation. Each evaluation word was rated on a 7-point scale ranging from "strongly agree" to "strongly disagree" and assigned a score from -3 to 3. These scores were used as impression values.

General Patterns. For the input images and impressions, 2,878 general pattern images (Fig. 2b) and evaluation words were selected and quantified as in the case of floral patterns.

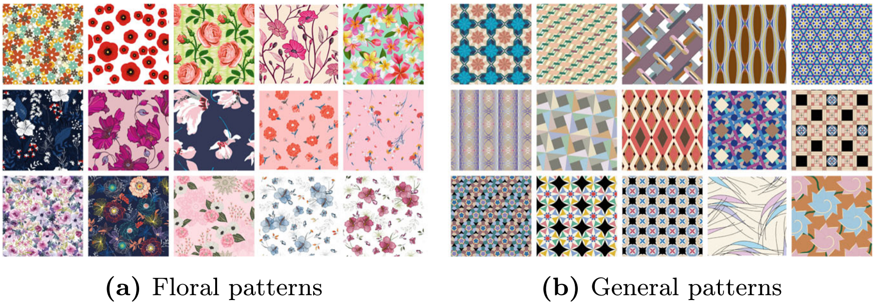


Fig. 2 Input pattern images

4.2 Implementation Details

We employ the Stable-Diffusion-v1-5 model. The model is trained with a batch size of 4, a learning rate of 10^{-6}, and for 10,000 steps. We train the proposed model with the prompt "a photo of S^*" as an input text. During inference, we generate images with three prompts: "a photo of S^*", "S^*", "S^* texture". For the general pattern dataset, we use an additional prompt of "floral texture of S^*".

4.3 Generated Results and Discussion

Figure 3a, b show examples of the generated images using floral patterns and general patterns, respectively, using the impression values of the original images and the labeled prompts as the inputs.

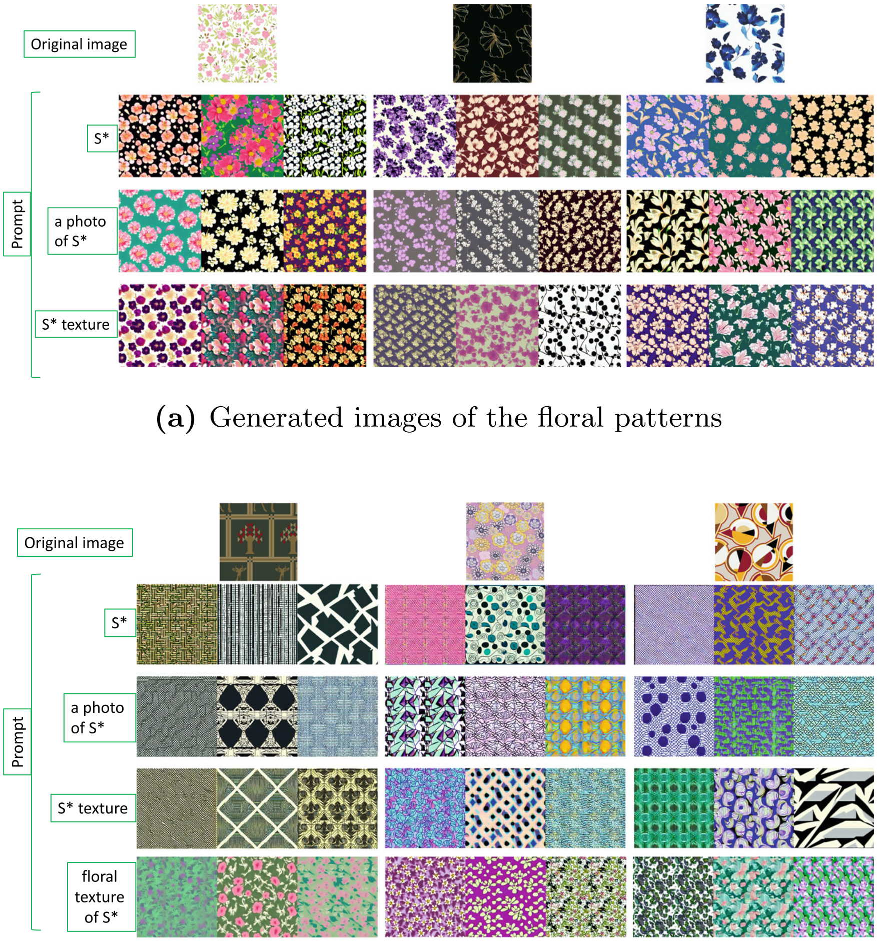
Discussion. When comparing the generated images from the three different prompts: "a photo of S^*", "S^*", and "S^* texture", we find that distinguishing differences between them is challenging. Thus, to validate whether the input text adequately conditioned the model, we utilized the prompt "floral texture of S^*". The successful generation of accurate floral patterns confirmed that the input text effectively conditioned the model.

Next, we evaluated whether the input impressions effectively conditioned the model. However, it was similarly challenging to discern noticeable differences. Therefore, we conducted an experiment to address this issue.

5 Verification Experiment

The proposed model is considered valid if it can be demonstrated that the generated image from the original image with a high evaluation for a specific impression also elicits a high impression, while an image generated from the original image with a low evaluation also elicits a low impression. Furthermore, if the evaluations of both the original and generated images are consistent, the model can be considered more accurate.

We conducted an experiment in which 30 participants rated their subjective impressions of the original and generated images using a 7-point scale. The ratings were then quantified on a scale of -3 to 3. We calculated both the correlation and the difference between the impression scores of the original and generated images. Through the following four steps, we selected appropriate images and evaluation words for the experiment.



(a) Generated images of the floral patterns

(b) Generated images of the general patterns

Fig. 3 Generated images

5.1 Selection of Dataset

To carry out the experiment, we selected a dataset from the general and floral pattern datasets. We first randomly chose five original images from each dataset. We then used the impressions of the original images to generate 15 images from each of the three prompts: "a photo of S*", "S*", and "S* texture". We generated a total of 225 images for each dataset and counted how many of the generated images were usable.

Usable images are defined as those without any noise and with well-established pattern structures. Examples of usable and unusable images are shown in Fig. 4.

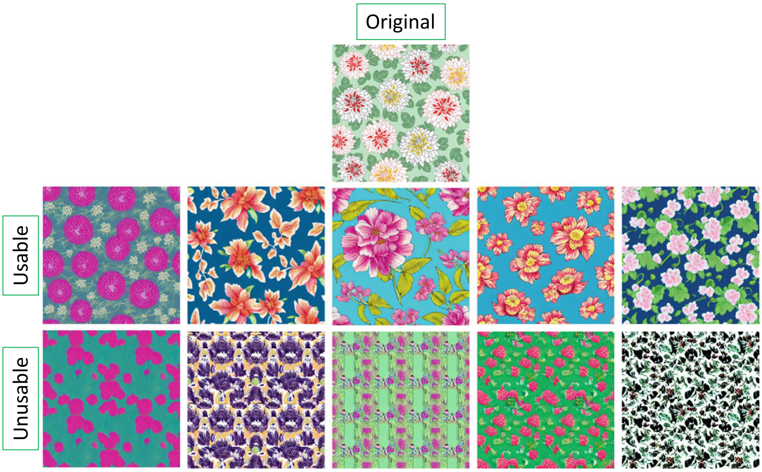


Fig. 4 Example of usable and unusable images

Table 1 Usable generated images

	Floral pattern	General pattern
a photo of S^*	28	13
S^*	27	10
S^* texture	20	9
Usable images/total images	75/225	32/225
Usable images percentage	33%	14%

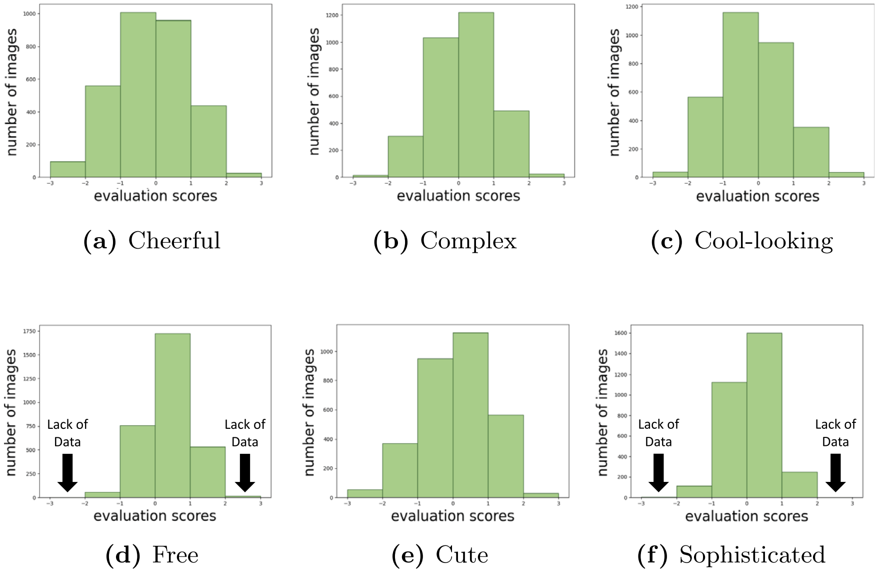
We confirmed that the number of usable images generated from the floral pattern dataset is larger than that of the general pattern dataset, as described in Table 1. Therefore, we used the floral pattern dataset to conduct the experiment.

5.2 Selection of Evaluation Words

Next, we selected the evaluation words from the 10 words of the floral pattern dataset mentioned in section "Floral Patterns". Prior research [4] has shown that these words are categorized into six factors (Table 2). We selected the word with the highest factor loading for each factor. However, for the evaluation words "free" and "sophisticated", the distribution of evaluation scores shows a noticeable lack of images assigned the highest ($+2 \sim +3$) and lowest scores ($-2 \sim -3$), indicating a limited representation of extreme evaluations (Fig. 5). Thus, excluding these two words, we selected four words (Table 3).

Table 2 Factors of evaluation words

Factor	Evaluation word
Pop	Cheerful Bright Colorful
Elaborate	Complex Multilayered
Refreshing	Cool-looking
Novel	Free
Tidy	Cute
Stylish	Elegant Sophisticated



(a) Cheerful (b) Complex (c) Cool-looking

(d) Free (e) Cute (f) Sophisticated

Fig. 5 Distribution of the evaluation word

Table 3 Selected evaluation words

Evaluation word
Cheerful
Complex
Cool-looking
Cute

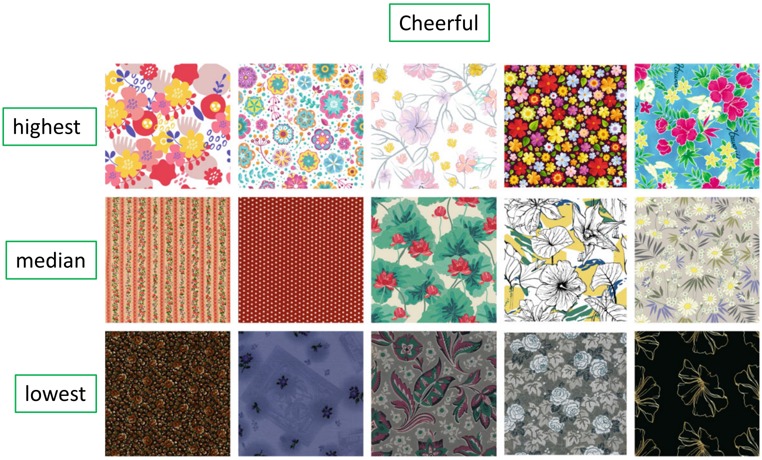


Fig. 6 Selected original images of the impression "cheerful"

5.3 Selection of Original Images

For each of the evaluation words, five images corresponding to the highest, median, and lowest evaluation scores were selected as shown in Fig. 6. This resulted in a total of 60 original images (4 evaluation words × 5 images × 3 score patterns = 60) being selected.

5.4 Selection of Generated Images

After generating images using the impression scores of the selected original images, we chose the generated images. First, we excluded images that are not usable as mentioned in Fig. 4.

Second, we used the DISTS metric [10] to measure similarity between the original and generated images. DISTS [10] is a full-reference image quality assessment (IQA) method that evaluates image similarity by combining structural and textural information. It is designed to align closely with human perceptual judgments of image quality. The lower the DISTS scores, the more objectively similar the images.

After calculating the DISTS scores, we selected three generated images with the highest, median, and lowest scores. By leveraging DISTS scores, we avoided intentional or subjective choices, ensuring an unbiased selection of the generated images.

Through this process (Fig. 7), we were able to select 600 images (60 original images × (1 original image + (3 prompts × 3 generated images based on DISTS))).

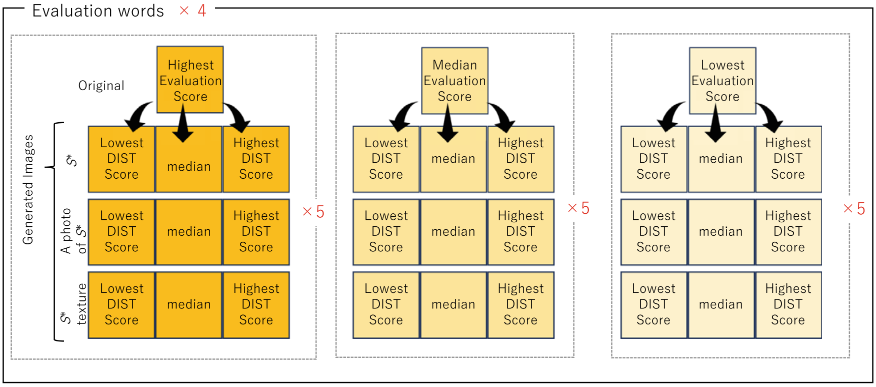


Fig. 7 Selection of images

5.5 *Results and Discussion*

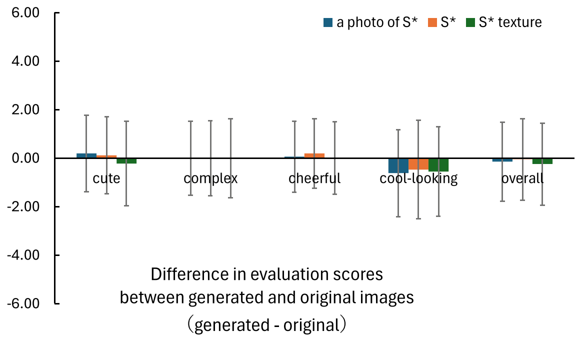
Correlation. Using the images from Sect. 5, we conducted a subjective evaluation experiment and calculated the correlation between the evaluation scores of the original and generated images for each participant, as shown in Table 4. Since the average correlation among participants was moderate, it can be said that the generated images show trends similar to the impressions of the original images. Regarding the prompts, although there is not much difference between the correlation coefficients of the prompts, the correlation coefficient of "a photo of $S*$" is slightly greater than the others. This indicates that the generated images from the prompt "a photo of $S*$" reflect the impressions of the original images. This may be because the prompt "a photo of $S*$" was used as the input text to train the model.

Regarding the DISTS scores, we observed that lower DISTS scores correspond to higher correlation coefficients. This shows that greater similarity between the original and generated images results in the generated image better reflecting the impressions of the original image.

Table 4 Average correlation coefficient between the evaluation scores of the original and generated images

Prompt	DISTS			Average	SD
	Lowest	Median	Highest		
$S*$	0.580	0.510	0.506	0.532	0.068
$S*$ texture	0.617	0.490	0.431	0.512	0.070
A photo of $S*$	0.622	0.565	0.453	**0.547**	0.065
Average	**0.606**	0.522	0.463		
SD	0.066	0.075	0.058		

Fig. 8 Difference in evaluation scores between generated and original images



Difference in Evaluation Scores. We calculated the difference in evaluation scores between the original and generated images, with 0 indicating no difference and the maximum difference being 6 or −6 as shown in Fig. 8. The average score difference for each participant was calculated, and the results showed minimal differences, with rounding often resulting in a score of 0. The evaluation word "cool-looking" had slightly lower scores for generated images, but the difference was less than 1. Overall, the results indicate a strong alignment between the original and generated images, demonstrating good performance.

Limitation. Although our method performed well in generating impression-reflective images, it still has some limitations. The success rate for generating patterned images is relatively low, with 33% achieved for the floral pattern dataset and only 14% for the general pattern dataset, as shown in Table 1.

6 Conclusion

In this study, we proposed a pattern image generation method in which impression scores can be inserted as input to condition image generation alongside input text. By generating images from the prompt "floral texture of S^*", we confirmed that the input text effectively conditions the model. We also verified that the model effectively reflects the input impressions by analyzing the correlation and differences between the subjective impression ratings of the original and generated images. From these results, we find that the prompt used to train the model reflects the impressions better. We also find that greater similarity between the original image and the generated image leads to closer alignment in their evoked impressions. Furthermore, the fact that generated images with high DISTS scores also correspond to high correlation coefficients confirms that our model generates objectively different images from the original images while preserving their original impressions. Lastly, we confirmed that the difference in the evaluation scores between the original and generated images is little to none, indicating the strong alignment of the impression between the images.

In the future, we will focus on improving the success rate of generating patterned images by expanding and diversifying the dataset. The current low success rate in generating patterned images is likely due to the relatively small size and limited variety of existing datasets. By collecting and training on larger datasets that encompass a broader range of patterns and designs, we aim to enhance the model's ability to generate consistent and high-quality outputs.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number JP23K24936.

References

1. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings Of The IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695
2. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with CLIP Latents, vol 1, p 3. arXiv:2204.06125
3. Saharia C et al (2022) Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in neural information processing systems, vol 35, pp 36479–36494
4. Sunda N, Tobitani K, Tani I, Tani Y, Nagata N, Morita N (2020) Impression estimation model for clothing patterns using neural style features. In: Stephanidis C, Antona M (eds) HCI international 2020 - posters. HCII 2020. Communications in computer and information science, vol 1226. Springer, Cham, pp 689–697.https://doi.org/10.1007/978-3-030-50732-9_88
5. Sugiyama Y, Sunda N, Tobitani K, Nagata N (2023) Texture synthesis based on aesthetic texture perception using CNN style and content features. In: Na I, Irie G (eds) Frontiers of computer vision. IW-FCV 2023. Communications in computer and information science, vol 1857. Springer, Singapore, pp 107–121. https://doi.org/10.1007/978-981-99-4914-4_9
6. Gal R, Arar M, Atzmon Y, Bermano A, Chechik G, Cohen-Or D (2023) Encoder-based domain tuning for fast personalization of text-to-image models. ACM Trans Graph (TOG) 42:1–13. https://doi.org/10.1145/359213
7. Shiohara K, Yamasaki T (2014) Face2Diffusion for fast and editable face personalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6850–6859
8. Julesz B (1981) Textons, the elements of texture perception, and their interactions. Nature 290(5802):91–97. https://doi.org/10.1038/290091a0
9. Portilla J, Simoncelli EP (2000) A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vis 40:49–70. https://doi.org/10.1023/A:1026553619983
10. Ding K, Ma K, Wang S, Simoncelli E (2022) Image quality assessment: Unifying structure and texture similarity. IEEE Trans Pattern Anal Mach Intell 44(5):2567–2581