一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IMQ2024-79,IE2024-157,MVE2024-96(2025-03)

# Stable Diffusion を用いた印象駆動型柄生成

ジェエントイサンジャ† 塩原 楓†† 山崎 俊彦†† 都賀美有紀† 飛谷 謙介††† 長田 典子†

† 関西学院大学大学院理工学研究科　〒669–1330 兵庫県三田市学園上ヶ原 1 番
†† 東京大学大学院情報理工学系研究科　〒113–8656 東京都文京区本郷 7–3–1
††† 情報科学芸術大学院大学メディア表現研究科　〒503–0006 岐阜県大垣市加賀野 4 丁目 1 番地 7
E-mail: †{sannjajn28,toga.m,nagata}@kwansei.ac.jp, ††{shiohara,yamasaki}@cvm.t.u-tokyo.ac.jp, †††tobitani@iamas.ac.jp

**あらまし**　本研究では印象に基づく柄生成技術を提案する. 具体的には, テキストプロンプトと柄画像から定量化された印象値と柄画像を学習データとし, テキストプロンプトと印象値を入力とする, Stable Diffusion モデルを用いた画像生成モデルを提案する. 提案手法により生成した画像とオリジナル画像に対して印象評価実験を行い, オリジナル画像と生成画像の相関関係と評価差を調査した. その結果, 生成画像が入力印象値を正確に再現できていることが確認された.

**キーワード**　感性, 印象, Stable Diffusion, 画像生成

# Aesthetic-Driven Pattern Design via Stable Diffusion

J N HTOI SANN JA†, Kaede SHIOHARA††, Toshihiko YAMASAKI††, Miyuki TOGA†, Kensuke TOBITANI†††, and Noriko NAGATA†

† School of Science and Technology, Kwansei Gakuin University
1 Uegahara, Gakuen, Sanda-shi, Hyogo, 669–1330 Japan
†† Department of Information and Communication Engineering, The University of Tokyo
7–3–1, Bunkyo, Tokyo, 113–8656 Japan
††† Information Engineering, Institute of Advanced Media Arts and Sciences
4–1–7 Kagano, Ogaki-shi, Gifu, 503–0006 Japan
E-mail: †{sannjajn28,toga.m,nagata}@kwansei.ac.jp, ††{shiohara,yamasaki}@cvm.t.u-tokyo.ac.jp, †††tobitani@iamas.ac.jp

**Abstract**　In this study, we propose an image generation model that generates images based on visual impressions. We used text prompts, visual impressions quantified from patterned images, and patterned images as training data. We then generate images from text prompts and quantified visual impressions using Stable Diffusion. To verify the effectiveness of our proposed method, we conducted a subjective evaluation. We then calculated the correlation between the evaluation scores of the original and generated images and the difference in evaluation scores. Through these results, we confirmed that our proposed method successfully reflects the input impressions.

**Key words**　Kansei, Visual impression, Stable Diffusion, Image generation

## 1. はじめに

プロダクトデザインの分野では, 個人の嗜好やライフスタイルを製品仕様に反映させるパーソナライゼーションへの要求が高まっている. 個人の価値観をプロダクトデザインに取り入れることによるユーザ体験の向上は, ユーザの満足度を向上させ, 個人および社会全体の Well-being を高めることに繋がる. 同時に, パーソナライズされた製品を提供することは, 不要な製品の生産を抑制し, 持続可能な消費に貢献できると期待できる. 特に, ファッション分野では, デザインの感性価値が重視されて

おり, パターンや色・質感が消費者の好みや感情に大きく影響を与えている．これらの要素を人工知能を活用してパーソナライズすることで, 消費者は専門的な知識を必要とせずに, 自分の好みに合った製品を容易に手に入れることが可能となる．

近年, 生成 AI, 特に Stable Diffusion [1], DALL-E2 [2], Imagen [3] などの text-to-image(T2I) 画像生成技術の急速な発展により, テキストプロンプトから所望の画像生成が可能となった．しかし, 現在のところ「かわいい」や「涼しげな」といった主観的な印象を表現する言葉から, 実際にその印象を持つパターン（柄）画像を生成する技術はまだ確立されていない．この課題に対処するため, 主観的な印象の定量化, 自動印象推定モデルに関する研究がされている [4]〜[6]．しかし, 印象を制御可能なパラメータとして画像を生成する手法には検討の余地がある．

そこで本研究では, Stable Diffusion (以下, SD) モデルを用いて, 印象に基づいた柄画像を生成する手法を提案する．

## 2. 先 行 研 究

### 2.1 感性的質感に基づく柄生成

パターンや柄に関連する研究として，テクスチャ（質感）解析の分野があり，これまでにさまざまなテクスチャ特徴量が提案されてきた [7]．しかし柄の印象のような，好き嫌いや良し悪しに関わる価値付けを伴う質感は，とくに感性的質感と呼ばれて区別されているが，これらを説明する特徴量については十分にわかっていない．寸田らは衣服の柄画像を表現する物理特徴としてスタイル特徴を用い，視覚的印象を自動で推定するモデルを構築した [4]．構築したモデルに基づき未知画像に対する感性的質感を推定したところ，人が実際に感じる感性的質感と強い正の相関があることが確認された．また, 谷口らの研究 [6] では, 布柄から喚起される印象に関する定量的な指標を学習データとして利用し, StyleGAN を用いてテクスチャ生成モデルを構築した．これにより, 潜在空間探索を通じてスタイルの制御を行い, 印象の制御を実現した．しかし, 同時に学習可能な印象評価語の数に限度（4 語）があり, それ以上の評価語を組み合わせて学習すると, 生成される画像の多様性が失われた．

### 2.2 パーソナライズされた画像生成モデル

テキストから画像を生成する Text to Image (以下, T2I) モデル [1]〜[3] は, 画像空間または潜在空間においてノイズ除去を定期的に行うことで画像を生成する．これらのモデルは, 入力テキストプロンプトを CLIP [8] のような言語と画像のマルチモーダルモデルを用いてエンコードし, 対応する画像を生成する．しかし, 入力テキストだけの情報で特定の対象や人物のパーソナライズした画像を生成するには不十分である．

この課題を解決するために, T2I モデルをパーソナライズする様々な手法が提案された [9], [10]．E4T モデル [9] は, 2 段階のパーソナライズプロセスとして, (1) 埋め込み単語の事前学習と (2) ドメイン固有のデータセットを用いたアテンション層の重みオフセットの調整で構成されており, 最後に fine-tuning を行うことでモデルは新しい概念を学習し, 小規模のデータサンプルからパーソナライズされた画像を短時間で生成することを可能とした．また, Face2Diffusion モデル [10] では, 顔の特徴を入力テキストに face embedding $S^*$ として入れることで, 特定の顔を条件とした画像生成が可能になった．この手法はカメラアングルの自由度を持たせつつ、顔の特徴を維持し、表情の過剰適応を抑え、背景を含めたテキストの忠実性を向上させた．

そこで, 本研究では, パーソナライズされた T2I モデルを用いて, 4 つ以上の印象の評価語を組み合わせることが可能で, 望む印象を反映した新しい柄画像を生成することを目的する．

## 3. 提 案 手 法

本研究では, 定量化された印象値を CLIP のテキスト空間内に Embedding ($S^*$) としてエンコードし, テキストプロンプトと合わせることで, 画像を所望の印象で条件つけることを目指す．提案手法の概念は図 1の通りである．本手法は T2I モデル (3.1) に印象エンコーダ (3.2) を組み合わせることで構成されている．この印象エンコーダによって, 入力印象のベクトルが印象 Embedding($S^*$) に変換され印象の特徴が学習できるようになっている．

### 3.1 T2I モデル

T2I モデルとして, SD [1] を用いる．SD はテキストエンコーダと VAE エンコーダの二つのエンコーダから構成されている．VAE エンコーダ $\varepsilon$ は入力画像 ($x$) を潜在空間に圧縮し, 次元を減らしながら画像の特徴を保持する．入力テキストプロンプト ($p$) は, CLIP テキストエンコーダ ($\tau$) を用いて, Word Embeddings ($\tau(p)$) にエンコードされる．UNet ($\epsilon_\theta$) は, 時間ステップ $t$ において, ノイズ潜在特徴 ($z_t$) を漸進的に除去し, テキスト特徴 ($\tau(p)$) に沿って, 追加されたノイズ ($\epsilon$) を予測する．モデルは, 式 1に示すように, 平均二乗誤差損失関数 ($L_{ldm}$) を用いて, 実際に追加されたノイズと予測されたノイズの差 ( ε ) を最小化するように学習する．この学習過程を通して, SD はテキストと画像の間意味のある関係性を学習し, テキストプロンプトに沿った高品質な画像の生成を可能にする．

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

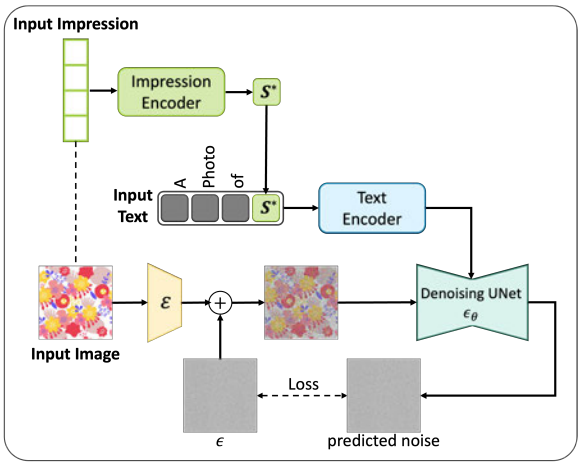
### 3.2 印象エンコーダ

印象エンコーダを用いて, 印象特徴を word embedding $S^*$ として抽出する．印象エンコーダ ($f_{imp}$) は, 入力画像 (x) に対応している印象値のベクトル ($v_x$) を 768 次元ベクトルに変換するニューラルネットワークである (式 2)．この変換は入力ベクトルを CLIP の埋め込みサイズに沿った高次元空間にマッピングする．ネットワークの各層には LeakyReLU 活性化関数を利用し, 印象ベクトルから意味のある特徴を効果的に抽出し, CLIP モデルとの統合を可能にする．

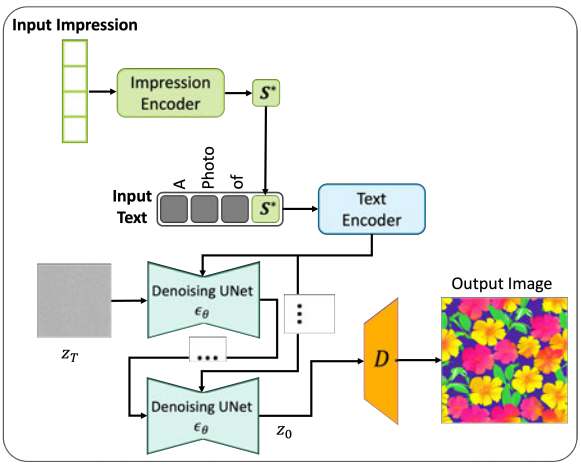
$$S^* = f_{imp}(v_x). \quad (2)$$

### 3.3 学習 (Training Process)

学習の際, (1) 印象値, (2) 画像と (3) テキストプロンプトを入力データとする．印象エンコーダは印象値のベクトルをエンコードし, 印象特徴 (embedding $S^*$) として抽出して, テキスト空間で処理する．embedding $S^*$ と入力テキストの組み合わせはテキストエンコーダによってエンコードされる．入力画像も




(a) 学習 (Training)




(b) 推論 (Inference)

図 1: 提案手法の概念図

表 1: 花柄の印象語 10 語

| | | | |
|---|---|---|---|
| かわいい | 明るい | 陽気な | 涼しげな |
| カラフルな | 自由な | 複雑な | 重なりのある |
| 上品な | 洗練された | | |

表 2: 一般柄の印象語 28 語

| | | | |
|---|---|---|---|
| かわいい | きれいな | やわらかい | ガチャガチャした |
| コントラストの高い | シックな | センスのある | 不揃いな |
| 冷たい | 凝った | 単色の | 古い |
| 古風な | 均等な | 対称的な | 平行な |
| 幾何学的な | 怪しげ | 集合体の | 高価な |
| 洋風な | 派手な | 特徴的な | 穏やかな |
| 細かい | 艶やかな | 華々しい | 規則的な |

VAE エンコーダ $\mathcal{E}$ によってエンコードされ, ガウスノイズ $\epsilon$ が付加される. 次に, すべての入力を UNet $\epsilon_\theta$ で処理し, 追加されたノイズ $\epsilon$ を予測する (図 1a).

### 3.4 推論 (Inference)

推論中入力データとして, (1) 印象値と (2) テキストプロンプトを用いる. 学習過程と同様に, 印象特徴量 $S^*$ を入力テキストプロンプトに組み込む. 次に, UNet によってノイズ除去を行うことで, 入力された印象とテキストを条件とした画像を生成する (図 1b).

## 4. 画像生成

### 4.1 データセット

(1) 花柄と (2) 一般的な柄の 2 つのデータセットを用いて画像生成を行った.

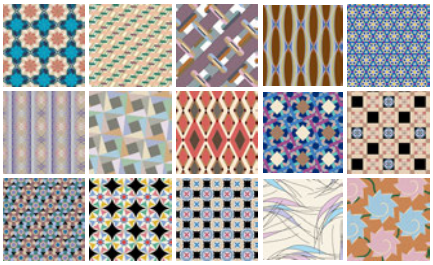
花柄データセットでは, 入力画像に対して, $512 \times 512$ サイズの花柄画像 3,098 枚を用意した (図 2a). 入力印象には, 花柄によって喚起される印象を記述した 10 個の評価語 (表 1) を用いた. これらの評価語は, 先行研究 [4] で行われた自由記述と適合度評価実験によって選択され, 主観評価によって定量化された. 主観評価では, 各評価語は「非常にそう思う」から「非常にそう思わない」までの 7 段階で評価され,–3〜3 点のスコアが割り当てられた. これらのスコアを印象値とする.

一般柄データセットでは, 入力画像として, $512 \times 512$ サイズの 2,878 枚の一般的柄画像 (図 2b) と入力印象値として 28 個の評価語 (表 2) を花柄と同様に定量化した値を使用する.

### 4.2 実装の詳細

Stable-Diffusion-v1-5 モデルを使用し, バッチサイズ 4, 学習率 $10^{-6}$, 10,000 ステップで学習を行った. “a photo of $S^*$” というプロンプトを入力テキストとして, 提案モデルを学習する. 推論時には, 3 つのプロンプト “a photo of $S^*$”, “$S^*$”, “$S^*$ texture” で画像を生成する. 般的柄のデータセットの際は,“floral texture of $S^*$” というプロンプトを追加し, 画像生成を行った.

(a) 花柄



(b) 一般柄

図 2: 柄画像の例

### 4.3 画像生成結果・考察

図 3a, 図 3bは花柄と一般柄のデータセットを用いて生成し


(a) 花柄の生成結果


(b) 一般柄の生成結果

図 3: 生成画像の例

た画像の例である．花柄と一般柄からオリジナル画像を選び，そのオリジナル画像の印象値と 3 つのテキストプロンプトを入力データとして画像生成を行った．

“a photo of $S^*$”, “$S^*$” と “$S^*$ texture” の 3 種類のプロンプトから生成された画像を比較すると，それらの違いを区別することは困難であることがわかる．そこで，モデルが 入力テキストによって適切に条件付けられているかを検証するために，プロンプト ‘floral texture of $S^*$” を用いて一般柄画像を生成した．正確な花柄が生成されていることから，入力テキストがモデルを効果的に条件付けていることが確認された．

次に，入力印象による効果を確認したところ，目立った違いを見分けることが困難であった．そこで，この課題を解決するため印象評価による検証実験を行った．

## 5. 検 証 実 験

提案モデルの妥当性を検証するため，印象評価実験を行った．仮説として，提案モデルが有効であれば，オリジナル画像において高い値の付いた印象はその画像をもとに生成した画像でも高く，オリジナル画像において低い値の印象は生成画像でも低くなっていると考えられる．さらに，オリジナル画像と生成画像の評価値に差がなく，一致していれば，より正確なモデルとみなすことができる．

実験では 30 名の被験者がオリジナル画像と生成された画像に対する主観的な印象を 7 段階で評価した．不誠実回答者を除く有効回答者数は 24 名であった．次に，評価を −3〜3 のスケールで定量化した．(1) オリジナル画像と生成画像の印象スコアの相関と (2) 個人ごとのオリジナルと生成画像との評価値の差を計算した．5. 1〜 5. 4の 4 つのステップを経て，実験に適した画像と評価語を選択した．

表 3: 使用可能な生成画像の比較

| | 花柄 | 一般柄 |
|---|---|---|
| a photo of $S^*$ | 28 | 13 |
| $S^*$ | 27 | 10 |
| $S^*$ texture | 20 | 9 |
| 使用可な画像枚数 | 75 / 225 | 32 / 225 |
| 使用可能な画像率 | 33% | 14% |

### 5. 1 データセットの選択

印象評価実験を行うために，花柄と一般柄のデータセットから一つのデータセットを選ぶ．最初に，各データセットからランダム 5 枚のオリジナル画像を選択し，そのオリジナル画像の印象を用いて 3 つのプロンプト “a photo of $S^*$”, “$S^*$” と “$S^*$ texture” からそれぞれ 15 枚の画像を生成した．各データセットについて合計 225 枚の画像を生成し，生成された画像のうち使用可能な生成画像枚数を調査した．使用可能な画像とは，ノイズがなく，パターン構造が確立されている画像と定義する．使用可能な画像使用不可能な画像の例を図 4に示す．表 3に示すように，花柄データセットから生成される使用可能な画像数は，一般柄のデータセットよりも多いことが確認された．そのため，花柄データセットを用いて実験を行う．

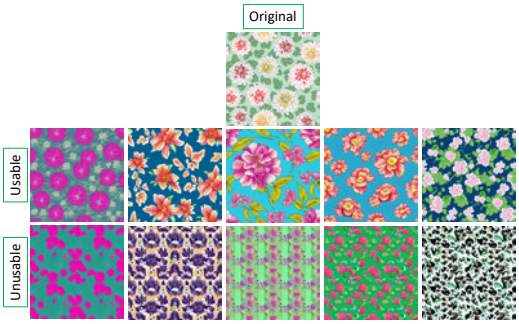



図 4: 使用可能な生成画像の例

表 4: 評価語を構成する 6 因子

| 因子 | 評価語 | | |
|---|---|---|---|
| ポップ | **陽気な** | 明るい | カラフルな |
| 凝った | **複雑な** | 重なりのある | |
| 爽やか | **涼しげな** | | |
| 斬新 | 自由な | | |
| 清楚 | **かわいい** | | |
| スタイリッシュ | 洗練された | 上品な | |

### 5.2 評価語の選択

花柄データセットの 10 語の評価語から使用する評価語を選定する．先行研究 [4] では，これらの評価語は 6 つの因子から構成されることが示された (表 4)．そこで各因子について，因子負荷量が最も高い評価語を選択した．ただし，「自由な」と「洗練された」は 最高スコア (+2 ～ +3) と最低スコア ( −2 ～ −3) で回答された画像がほとんどなく，中程度の評価に分布が限られていた．従って，この 2 つの評価語を除いた 4 つの評価語（陽気な，複雑な，涼しげな，かわいい）を選択した．

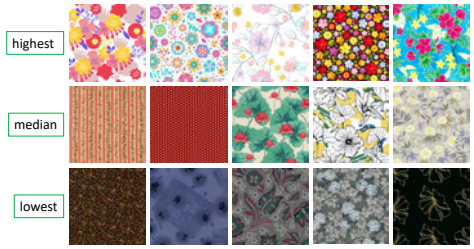
### 5.3 オリジナル画像の選択

各評価語について，図 5に示すように，評価値が最も高い，中程度，最も低い画像を 5 枚ずつ選定した．その結果，合計 60 枚のオリジナル画像 (4 評価語 × 5 枚の画像 × 3 高中低のパターン = 60) が選択された．

### 5.4 生成画像の選択

選択されたオリジナル画像の印象値を用いて画像を生成した後，評価実験に使用する生成画像を選定する．まず，図 4で述べたように，使用できない画像は除外した．次に，DISTS メトリック [11] を用いて，オリジナル画像と生成画像の類似度を測定した．DISTS は，構造情報とテクスチャ情報を組み合わせて画像の類似性を評価する全参照画質評価 (IQA) 手法で，人間の画質に関する知覚判断と近いことが示されている．DISTS スコアが低いほど，客観的に類似した画像であることを示す．

DISTS スコアを計算した後，スコアが最も高い，中程度，最も低いの 3 つの生成画像を選択した．DISTS スコアを活用して意図的でない選択をすることで，生成された画像について主観的な偏りのない選択を行った．以上の手順により (図 6)，600 枚の画像 (60 オリジナル画像 × (1 オリジナル画像 + (3 プロンプト × 3 DISTS による生成画像))) を選択することができた．




(a) 陽気な


(b) 複雑な


(c) 涼しげな


(d) かわいい

図 5: 各評価語の選択されたオリジナル画像

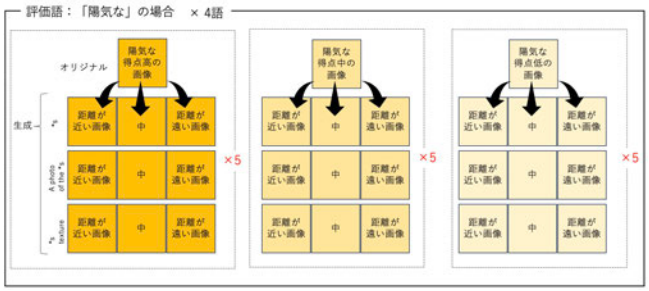



図 6: 画像選択の説明図

### 5.5 結果・考察

以上で選定したオリジナルと生成柄画像を用いて主観評価実験を行い，表 5 に示すように各参加者のオリジナル画像と生成画像の評価値の相関を算出した．参加者間の平均相関は正の相関であることから，生成画像はオリジナル画像の印象と同様の傾向を示していることがわかる．

プロンプトについて，異なるプロンプト間の相関係数には著

表 5: オリジナルと生成画像の評価値の平均相関

| 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 | | |

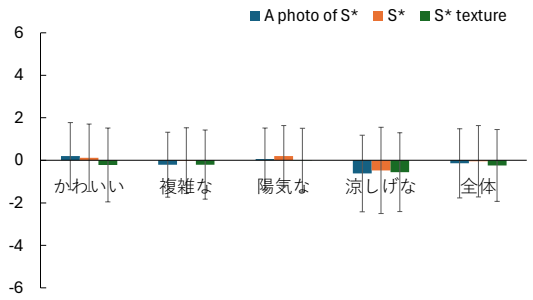



図 7: 生成画像とオリジナル画像との評価差

しい差がないものの, “a photo of $S^*$” の相関係数が他より少々高くなっており, プロンプト “a photo of $S^*$” から生成された画像がオリジナル画像の印象をより反映していることを示している. この原因として, プロンプト “a photo of $S^*$” を入力テキストとしてモデルを学習させたためと考えられる.

DISTS スコアについては, DISTS スコアが低いほど相関係数が高いことが確認された. これは, オリジナル画像と生成画像の客観的類似度が高いほど, 生成画像がオリジナル画像の主観的な印象をよりよく反映することを示している.

実験参加者ごとにオリジナル画像と生成画像の間の評価値の差を計算し, 可視化した（図 7）. 0 は差がないことを示す. 図中の差の値はどれも小さく, 「涼しげな」以外は四捨五入して 0 であった. 「涼しげな」は他の評価語と比べて差は大きかったがすべて 1 未満であった. 全体として, オリジナル画像と生成画像の印象の評価値は近く, 良い結果が得られたと考えられる.

### 5.6 課　　題

提案手法は, 先行研究 [6] の課題点を解決し, 4 つ以上の印象値の入力が可能で, 印象を反映した画像を生成するのに有効であるが, 柄画像を生成する成功率が比較的低いという課題がある. 表 3に示すように, 花柄データセットでは 33%, 一般柄データセットでは 14% に留まっており, さらなる工夫が必要である.

## 6. 結　　論

本研究では, テキストと共に印象値を条件として入力する柄画像生成手法を提案した. “floral texture of $S^*$” というプロンプトから画像を生成することで, 入力テキストがモデルを効果的に条件付けることを確認した. また, オリジナル画像と生成画像の主観的な印象評価の相関や評価差を分析することで, モデルが入力印象を効果的に反映することが示され, モデルの学習に使用したプロンプト “a photo of $S^*$” は, 印象をよりよく反映することが示された. さらに, DIST スコアが低い画像, つまりオリジナル画像と生成画像との客観的な類似性が高いほど, 喚起される印象が近くなることがわかった. 加えて, DISTS スコアが高い生成画像も相関係数が高いことから, 提案手法は元の印象を保持したまま, 客観的に異なる画像を生成していることが確認された. 最後に, オリジナル画像と生成画像の間の評価スコアの差はほとんどないことが確認され, 画像間の印象の強い一致が示された.

今後は, 課題点である画像の生成成功率を向上させることに注力する. 柄画像を生成する成功率が低いのは, 既存のデータセットのサイズが比較的小さく, 種類が限られているためと思われる. より広範なパターンとデザインを包含する, より大規模なデータセットを収集し, 学習することで, 一貫性のある高品質な出力を生成するモデルの能力を向上させることを目指す.

## 文　　献


[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10684-10695, 2022.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, “Hierarchical text-conditional image generation with clip latents,” ArXiv Preprint ArXiv:2204.06125, 2022.

[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, et al., “Photorealistic text-to-image diffusion models with deep language understanding,” Advances In Neural Information Processing Systems, vol.35, pp.36479-36494, 2022.

[4] N. Sunda, K. Tobitani, I. Tani, Y. Tani, N. Nagata, and N. Morita, “Impression estimation model for clothing patterns using neural style features,” HCI International vol.1226, pp.689-697, 2020. DOI:10.1007/978-3-030-50732-9_88

[5] Y. Sugiyama, N. Sunda, K. Tobitani, N. Nagata, “Texture synthesis based on aesthetic texture perception using CNN style and content features,” Frontiers of Computer Vision (IW-FCV), vol.1857, pp.107-121, 2023.

[6] 谷口史果, 飛谷謙介, 都賀美有紀, 長田典子, “Textile-GAN – StyleGAN の潜在空間探索による印象制御に基づくテクスチャ柄生成 –,” 信学技報, vol.123, no.433, pp.144-149, 2024.

[7] J. Portilla, E.P. Simoncelli, “A parametric texture model based on joint statistics of complex wavelet coefficients,” International Journal of Computer Vision, vol.40, pp.49–70,2000.

[8] A. Radford, et al., “Learning transferable visual models from natural language supervision,” International Conference on Machine Learning, pp.8748-8763, 2021.

[9] R. Gal, M. Arar, Y. Atzmon, A. Bermano, G. Chechik, D. Cohen-Or, “Encoder-based domain tuning for fast personalization of text-to-image models,” ACM Transactions On Graphics (TOG), vol.42, pp.1-13, 2023. DOI:10.1145/359213

[10] K. Shiohara, T. Yamasaki, “Face2Diffusion for Fast and Editable Face Personalization,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.6850-6859, 2024.

[11] K. Ding, K. Ma, S. Wang, & E. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” IEEE Transactions On Pattern Analysis And Machine Intelligence, vol.44, no.5, pp.2567-2581, 2022.