ノードとスライダで細部編集を追い込む画像生成システムの提案と評価

大友千宙[1]　宮下芳明[1]

概要：Text-to-Image モデルで生成した画像を意図した通りの画像となるよう追い込むことは難しい．生成される画像をテキストのみで制御することが困難なためである．入力するテキストは単語の順序や修飾関係を明確にしたものでなければならず，後からの修正も容易ではない．また，生成された画像をテキストのみで編集する手法が研究されているが，スクリプト言語でのコーディングが必要となる．本研究はパラメータの調整や特定の部分のみを編集することによってユーザ自らが求める表現へと追い込めるようにすることを目的とし，テキストボックスとスライダからなるノードの操作によって画像の生成と編集を行うシステムを提案する．単語の入れ替えや修飾関係の指定，パラメータの調整といった作業はテキストボックスやコードエディタで行うのではなく，それに適したインタフェースで行うべきだと考えている．ノードベースのシステムとすることで単語同士の関係を指定することや，編集したい部分のみ操作することが可能になる．また，各ノードに備えたスライダ操作によってインタラクティブにパラメータの調整ができる．さらに，作例制作を通した提案システムの評価を行った．提案システムによって生成済みの画像を追い込むことでユーザが求める表現を持つ画像を制作できるようになることを示した．

1. はじめに

Stable Diffusion [1] などの Text-to-Image モデルによる画像生成では，意図した通りの画像を生成することは難しい．プロンプトと呼ばれるテキストのみで画像を制御するため，様々な要件を満たしたプロンプトを作成しなければならない．例えば，プロンプト内における単語の順序は生成に大きな影響を及ぼすため，生成したい画像に合わせて単語の順序を考慮する必要がある．また，単語の係り受け関係が不明確だと色などの属性が誤った個所に適用されてしまう．単語を一つ入れ替えたり並び替えたりする度にこうした点について再度考慮する必要がある．

ユーザが意図した通りの内容をテキストで表現しても，生成される画像はテキストの通りにならない．実際に，Stable Diffusion がプロンプト内の単語を無視した画像を生成することが指摘されている [2]．この問題への対処として，プロンプトのみで生成される画像を正確に制御する手法が研究されている．これにより，特定部分のみの編集や単語の重みづけの変更による画像の編集が可能となる．しかし，こうした編集手法は作業内容に応じてスクリプト言語によるコーディングを要する．単語や数値を一つでも変更する度にコーディングをしなければならない．

本研究の目的は画像生成においてユーザ自身が求める表現を追い込めるようにすることである．ユーザ自らが表現を追い込めるようにするためには，単語の並び替えや係り受けの明確な指定，細部の編集といった作業に適したインタフェースが必要だと考えた．そこで，テキストボックスとスライダを有したノードでプロンプトを構築し，個別のノードを操作することで画像の生成と編集が行えるシステムを提案する．ノードベースのシステムとすることで，単語の入れ替えが容易になるとともに，明示的に単語同士の関係を入力することができる．更に，スライダを操作するだけでパラメータの調整が可能となる．提案システムではノードの操作に合わせて自動でプロンプトを生成しスクリプト言語によるコードを作成する．よって，ユーザは手間のかかるプロンプト作成やスクリプト言語でのコーディングをすることなく，表現を追い込むことに集中できる．

また，作例の制作による提案システムの評価を行った．提案システムによってユーザの意図した表現に近づけた画像や，編集手法の適用によってトレーニングデータへの偏りを抑えた画像を制作した．また，提案システムによってどこまで表現を追い込めるかを検証するため，インターネットに存在しない写真の再現画像を制作した．これらの作例を通した評価から，提案システムによってユーザが求める表現へと追い込めるようになることを示した．

[1] 明治大学

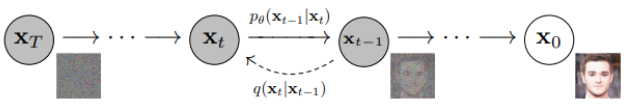

図 1 Diffusion model の概要．ノイズを付与する Forward process とその逆をたどる Reverse Process の組み合わせによってデータを生成する（ [3] より引用）．

図 2 Prompt-to-Prompt による画像編集の概要．cross-attention map の操作によって編集を実現している（ [4] より引用）．

2. 関連研究

2.1 テキストによる画像生成

テキストから画像を生成する Text-to-Image モデルには Stable Diffusion [1] を始めとして，Imagen [5] や DALLE-2 [6] などがある．本研究で利用した Stable Diffusion は，Diffusion model [3,7] を基にしている．Diffusion model は生成モデルの一つであり，生成品質が高く様々な用途に拡張できるという特徴がある．従来の生成モデルと比較すると，学習の安定性や動画などの生成が難しかったデータを生成できることなどが強みとして挙げられる．

Diffusion model は図 1 のように与えられた画像に少しずつノイズを加えていく Forward Process によって学習を行う．学習済みの Diffusion model においてノイズが付与された画像からノイズを推定し，少しづつ取り除く Reverse Process を行うことによってデータの生成が可能となる．だが，高解像度の画像を生成しようとすると計算量が膨大になるという問題がある．そこで，Stable Diffusion では Diffusion model の入出力の前後に Variational AutoEncoder（VAE）を利用し，高解像度の画像を低次元の画像として扱うことで効率の良い学習と画像生成を実現している．また，attention と呼ばれる機構を利用することでテキスト情報を用いた条件付け生成を行うことができる．

Text-to-Image モデルによる画像生成の特徴として，同じプロンプトやパラメータでも生成の度に構図や内容が変わることがある．この特徴により，ユーザは様々なバリエーションを手軽に生成することができる．だが，生成された画像の一部のみを変更することは難しい．プロンプトを一部でも変更すると全く異なる画像が生成されるためである．生成結果の固定は生成初期のノイズ画像を指定するシード値の固定により可能だが，シード値を変更しても現在の出力を踏まえた編集は難しい．

本研究での提案システムのように，ノードの操作による画像の生成が可能なシステムが既に存在しており，代表的なものに ComfyUI [8] がある．こうしたシステムは後述する画像生成のための編集手法に対応したものが少なく，ノードの操作で特定の個所を編集することができない．

2.2 テキストの編集による画像編集

Text-to-Image モデルによって生成される画像をテキスト操作のみで正確に制御，編集するための手法が研究されている．Hertz らは Stable Diffusion などのモデルではテキストを少し変更するだけで全く異なる画像が生成されることを指摘した．そこで，テキストによる素早く直観的な編集を妨げることのないよう，テキストのみによって編集を制御する Prompt-to-Prompt を提案した [4]．Prompt-to-Prompt によって図 2 のように単語の書き換えや単語の追加による画像の編集が可能になる．Stable Diffusion では cross-attention によってプロンプトの意味的内容が取り込まれ，生成する画像にその意味的内容が反映される．Hertz らは画像の構造と内容は入力されたテキストと入力画像から算出される値，cross-attention に依存することを発見し，テキストの編集に合わせて cross-attention を入れ替えたり，増減させることで画像の編集を可能にした．

また，Tumanyan らは入力した画像の構図を保ったままユーザが入力したテキストの内容に従う画像を生成することでユーザによる制御を実現する手法，Plug-and-Play を提案した [9]．Tumanyan らはテキストのみによる画像生成はユーザの制御性を犠牲にしていると指摘した．そこで，画像の入力によってレイアウトを制御し，テキストによってレイアウト内の内容や外観をガイドする Image-to-Image に焦点を当てた Plug-and-Play を提案している．Plug-and-Play は cross-attention の操作ではなく self-attention の操作によって画像の編集を可能にしている．self-attention はテキスト情報からではなく，入力画像のみから算出される値である．Tumanyan らは self-attention の操作によって Prompt-to-Prompt では保持されなかった局所的な空間情報を保持することに成功している．さらに，self-attention はテキスト情報を入力にとらないという特徴がある．そのため，Prompt-to-Prompt では必須だった編集前の画像を生成する際に利用したプロンプトを必要とせず，任意のテキストと画像のみで生成することが可能である．

この 2 つの手法を基にした画像編集手法も研究されており，Ge らはリッチテキストによって生成される画像の構図や内容を制御する手法を提案した [10]．Ge らはテキストによる画像生成はテキストベースのタスクであり，プレーンテキストしか使用できないことは問題であるとした．そこで，サイズや色などを変更可能なリッチテキストエディタによる画像編集手法を提案している．リッチテキストにより，ユーザはより正確に希望する画像を記述できる．

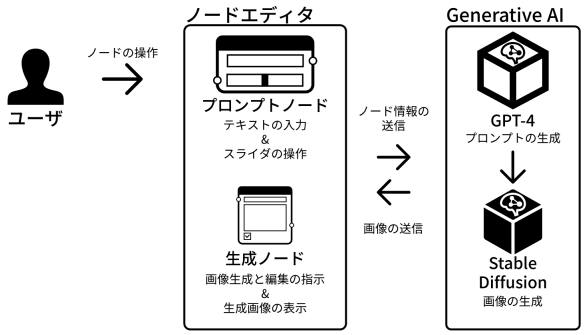

図 3 提案システムの概要．ノードエディタ内の操作に合わせてプロンプトや画像が生成される．

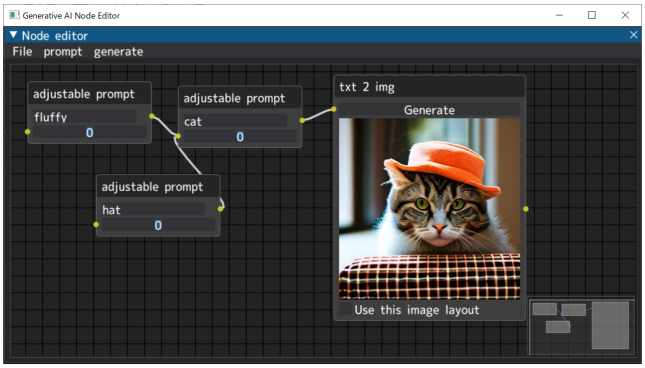

図 4 提案システムのスクリーンショット．画面全体がエディタであり，ノードの操作が可能である．画面右下にはノードの位置や接続状況が示されたミニマップが表示される．テキストが入力されているノードがプロンプトノード，生成した猫の画像が表示されているノードが生成ノードである．

ここで取り上げた 3 つの手法はテキストによる画像生成の制御を目的としている．そのため，本研究の目的に即したインタフェースやシステムを提案するものではない．

3. 提案システム

3.1 システム概要

提案システムはノード接続やスライダ操作により画像の生成と編集を行うノードベースシステムである．提案システムの概要とスクリーンショットをそれぞれ図 3，図 4 に示す．提案システムはユーザが操作するノードエディタ，画像生成とプロンプト生成を行う Generative AI によって構成される．ノードエディタは Python3.10 と Python のライブラリである DearpyGUI によって実装した．加えて，DearpyGUI によって作成された画像処理ツールである Image-Processing-Node-Editor[*1]のソースコードを一部利用した．また，画像生成モデルには Stability AI の stable-diffusion-2-1-base を使用した．加えて，モデルを使用するためのインタフェースとして，Hugging face が提供するライブラリの pipelines を使用した．画像生成用のプロンプト作成は Open AI の GPT-4 を利用した．

[*1] 高橋かずひと氏による画像処理アプリ
https://github.com/Kazuhito00/Image-Processing-Node-Editor

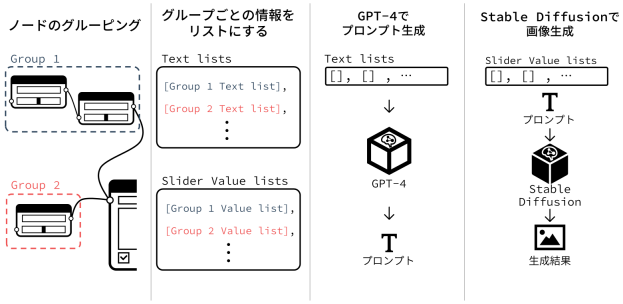

図 5 提案システムにおける画像生成時の動作

提案システムには横長のプロンプトノードと縦長の生成ノードが存在する．ユーザはノードエディタ内でプロンプトノードを配置し接続することでプロンプトを構築する．プロンプトノード同士の接続が完了した後に生成ノード内のボタンを押下することで画像を生成する．画像生成後にノードの追加，テキストの入れ替え，スライダによる調整といった操作により画像を編集する．

3.2 提案システムによる画像生成

画像生成における提案システムの動作を図 5 に示す．まず，提案システムはボタン押下時のノード接続情報を基にプロンプトノードをグルーピングする．具体的には，生成ノードに直接接続しているノードを親ノードとし，そのノードに接続しているノードを子ノードとしてまとめる．グルーピングによりノードの接続状況を反映したプロンプトを生成することができる．そして，グループごとにノードが持つ情報のリストを作成する．ノードが持つテキストボックス内のテキストとスライダの値のリストは，GPT-4 への入力や単語の意味的効果をどれほど変化させるかを決定する際に利用される．その後，テキストのリストを GPT-4 に入力しプロンプトを生成する．こうして得た画像生成用のプロンプトを Stable Diffusion に入力することで画像を生成する．

3.3 ノードとスライダによる画像の編集

提案システムではノードの追加，テキストの入れ替え，スライダの操作という 3 つの操作とこれらの操作の組み合わせにより画像編集ができる．本節では図 4 のノードの接続状況と生成画像を基に，3 つの操作によってどのように画像が編集されるかを示す．なお，ノードの追加とテキストの入れ替えによる生成結果の編集には Plug-and-Play，スライダによる調整は Prompt-to-Prompt の Attention Re-Weighting によって実装した．

3.3.1 ノードの追加

生成ノードに直接プロンプトノードを追加すると，画像

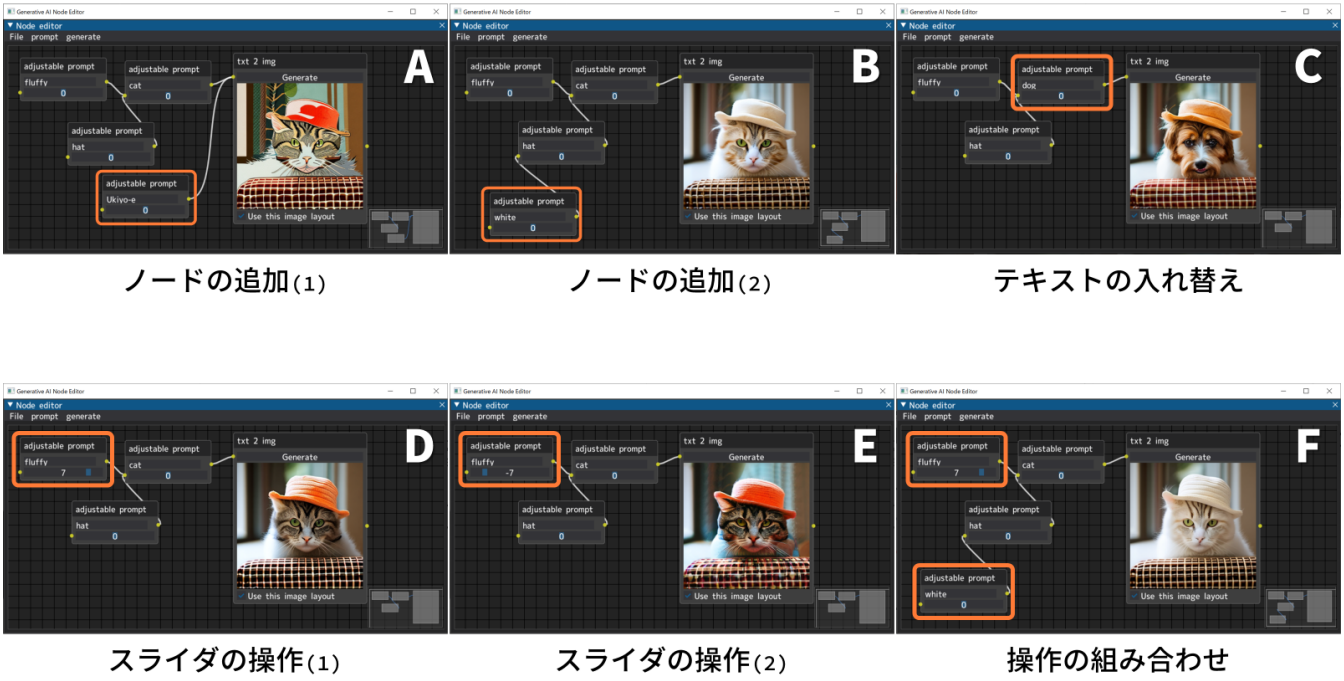

図 6 提案システムによって画像を生成・編集した際のスクリーンショット

のスタイル変換などの画像全体に対する編集が可能である．図 6A は生成ノードに直接「Ukiyo-e」とテキストボックスに入力したプロンプトノードを追加した際のスクリーンショットである．もともと生成されていた画像の構図や内容は維持したまま，画像全体が浮世絵のようなスタイルになったことがわかる．

また，既存のプロンプトノードに新たなプロンプトノードを接続すると特定のオブジェクトが持つ属性を指定することができる．図 6B は「hat」と入力しておいたノードの左に「white」と入力したノードを追加した際の様子である．「white」の後に接続されたノードである「hat」と「cat」に影響が及び，生成された画像では帽子と猫が白色になったことがわかる．また，生成ノードに直接プロンプトノードを接続した場合とは異なり，背景などの要素には影響が及ばない．

3.3.2 テキストの入れ替え

既存のプロンプトノードのテキストを直接書き換えるか，ノードごと入れ替えることで現在生成した画像の構図を維持したまま画像内のオブジェクトを入れ替えることができる．図 6C は既存プロンプトノードの「cat」を「dog」に書き換えて生成した際の様子である．元々の構図は維持されたまま，画像内に映っていた猫が犬になった画像が生成されたことがわかる．

3.3.3 スライダの操作

プロンプトノード下部のスライダを右方向に動かすことで，そのプロンプトノードが持つテキストの効果を強めることができる．図 6D はテキストボックスに「fluffy」と入力されているノードのスライダを右方向に操作した際の様子である．「fluffy」の意味的効果が増幅し，画像内における猫の毛量が増加している．

一方，スライダを左方向に動かすとテキストの効果を弱めることができる．図 6E は図 6D と同一のプロンプトノードにおいて左方向にスライダを動かした際の様子である．「fluffy」の意味的効果が減衰し，猫の毛量が減少していることが確認できる．

3.3.4 操作の組み合わせ

3 つの操作を組み合わせることで同時に複数の編集を重ねることが可能である．図 6F はテキストボックスに fluffy と入力したノードのスライダを右に動かし，「white」と入力したノードを追加した際の様子である．図 6B 内の画像のように猫と帽子が白色になり，図 6D 内の画像のように毛量が増加したことがわかる．

4. 作例

本節では提案システムによって制作した 3 つの作例について述べる．作例の制作は著者自らが行った．

4.1 作例 1：単語の強調による表現の追い込み

図 7A は「sunset landscape」をプロンプトとして提案システムによって生成した画像である．図 7A は全体的に白い光で照らされており，夕暮れの風景画とするには「夕暮れ」をより強く画像に反映させるための編集が必要だと考

図 7 提案システムによって「夕暮れの風景画」を生成し，編集した際の画像．「sunset」のスライダ操作によって光の色が赤色に近くなった．

えた．そこで，図 7 下に示したように「sunset」と書かれたノードのスライダを右に動かして図 7B へと修正した．図 7B は図 7A の構図や雰囲気を保ったまま全体が赤い光で照らされている．特に図 7B 下部の草が赤みがかったことにより，図 7A より「夕暮れ」が強調され，「夕暮れの風景画」らしくなったといえる．

編集によって「夕暮れ」らしくなったかを評価するため，CLIP スコアを算出した．CLIP [11] は OpenAI が公開したテキストと画像の類似度を算出することが可能な画像分類モデルである．本研究では CLIP モデルに日本語テキストに対応した clip-vit-b-32-japanese-v1 を使用した．また，本節では「sunset landscape」と，反対の意味を持つ「sunrise landscape」を対象のテキストとした．さらに，図 7A・B を対象の画像としてスコアを算出した．また，本節で算出した CLIP スコアは 1 枚の画像に対して複数のテキストとの類似度を算出する方式であり，1 つの画像における全テキストとのスコアを合算すると 1 になる．ここで，対象のテキストを「sunset landscape」のみにしてスコアを算出すると図 7A のスコアが 1，図 7B のスコアが 0 になった．本節でのスコア算出方式において一つのテキストのみを対象にして複数の画像とのスコアを算出すると，いずれかの画像のスコアが 1 となり，その他の画像はスコアが 0 になってしまう．本節で CLIP スコアを算出した意図は提案システムによってより「夕暮れ」らしくなったかを評価するためであり，どちらの画像が「夕暮れ」かを判定するためではない．よって，「sunset」と現象的には同一だが単語としては反対の意味を持つ「sunrise」を利用した「sunrise landscape」を対象のテキストに追加した．スコアを算出するテキストを 2 種類にすることでとあるテキストに対する画像のスコアが 0 となることを回避した．

算出したスコアから，編集後の画像の方が「sunset landscape」との類似度が低いことがわかった．図 7A と「sunset landscape」とのスコアは 0.94 であり，図 7B とのスコアよりも 0.12 高いスコアとなっている．Stable Diffusion は CLIP のテキストエンコーダを画像生成時に利用しており，スコアが高くなるように内部で生成を繰り返すことでテキストの内容に沿った画像を生成している．そのため，図 7A は Stable Diffusion が生成する画像としてはもっともらしいといえる．一方，図 7B と「夕暮れの風景画」とのスコアは 0.82 とやや低く，こうした画像を Stable Diffusion から直接得るためには何度も生成するか，CLIP の処理をスキップするパラメータをユーザ自身で調整するしかない．よって，既存の編集手法による単語の重み変更をスライダに結びつけたことでユーザ自らの表現を「追い込む」ことが可能となったと考えている．

4.2 作例 2：学習データと異なる画像の生成

図 8 は「van Gogh の夜のカフェテラス」を生成した際の提案システムのスクリーンショットである．生成された画像は，画像内左側に黄色で描かれたカフェテラスがあり，夜空は青色で描かれているという特徴を持つ．シード値を変更して生成しても左に黄色のカフェテラス，夜空が青で描かれた同じような画像が生成される．シード値を変更しても似たような画像が生成される理由として，Stable Diffusion が学習したデータの中に Vincent van Gogh の「夜のカフェテラス」が含まれていることが考えられる．Stable Diffusion は学習のためのデータセットとして LAION-5B [12] を利用している．その中には Vincent van Gogh が 1888 年に発表した「夜のカフェテラス」が含まれている．そのため，シード値を変更してもプロンプトが「van Gogh の夜のカフェテラス」である限りは似たような画像が生成されると考えた．よって，1888 年の「van Gogh の夜のカフェテラス」とは異なる構図を持つ「van Gogh の夜のカフェテラス」を Stable Diffusion で生成したければ，前節で述べた通り何度も生成するか，CLIP の処理をスキップするパラメータを調整する必要がある．また，Prompt-to-Prompt や Plug-and-Play といった編集手法を利用することで学習データから離れた画像を生成することができるが，これらの手法を利用するためにはプログラムを記述する必要がある．スライダで任意のパラメータを調整したり，ノードベースシステムのように個々のノードを操作するだけで編集手法を適用することはできない．

そこで，提案システムによって「夜のカフェテラス」の画像を生成し，後から van Gogh と入力したノードを追加すれば 1888 年の「夜のカフェテラス」とは異なる画像が得られると考えた．図 9A は「夜のカフェテラス」を生成した際の提案システムのスクリーンショットと生成された画像である．また，図 9B は「night」と入力したノードのスライダを右方向に操作した際の提案システムのスクリーンショットと生成された画像である．そして，図 9C は細部編集後に van Gogh ノードを追加した際の提案システムのスクリーンショットと生成された画像である．図 9C は 1888 年の「夜のカフェテラス」とは全く異なる構図であり

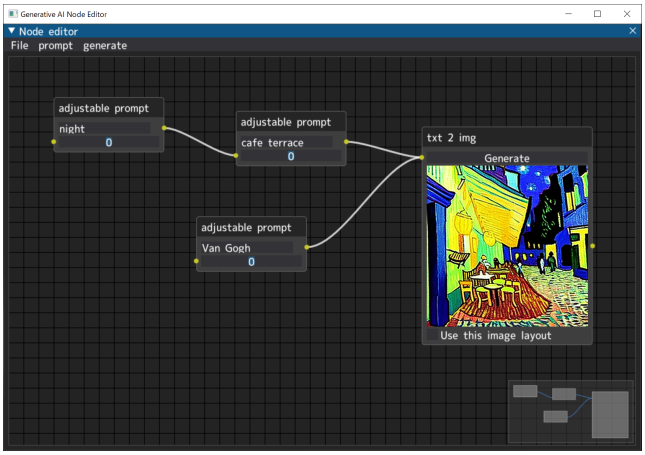

図 8 「van Gogh's cafe terrace at night」をプロンプトとして画像を生成した際の提案システムの様子．生成された画像は 1888 年に発表された「夜のカフェテラス」と同じ構図や雰囲気を持っている．

ながら Van Gogh の特徴的な青色と黄色が表れている．なお，図 9A として示した編集前の画像は「夜」とするには全体的に明るかったため，「night」と入力されているノードのスライダを右方向に動かした．こうして得られた画像が図 9B であり，この状態から van Gogh と入力したノードを追加することで図 9C を得た．

図 9C は Stable Diffusion から得ることが困難な画像であることを示すため，作例 1 と同様の手法で CLIP スコアを算出した．算出したスコアを図 10 に示す．対象とした画像は図 9A・B・C であり，対象としたテキストは「van Gogh's cafe terrace at night」と「cafe terrace at night」の 2 つである．図 10 から，図 9A・B・C と「van Gogh's cafe terrace at night」とのスコアが著しく低いことがわかる．特に図 9C とのスコアは 0.04 と対象にした画像の中でも最も低い．前述したように，Stable Diffusion はスコアが上昇するように内部で生成を繰り返すため，このような低いスコアの画像を Stable Diffusion のみで生成することは難しい．よって，提案システムを介して複数の異なる編集手法を同時に適用したことで，学習データから離れた画像の生成と編集が可能となり，ユーザ自身が求める表現へと追い込めたと著者は考えている．

4.3 作例 3：インターネットにない画像の再現

提案システムによって表現を追い込むことが可能であることを検証するため，著者が撮影した写真（図 11A）の再現を試みた．図 11A はインターネットにアップロードしていないため，この画像を学習データとしているデータセットは存在しない．図 11A は山と空，手前に雑木林があり，画像の半分ほどは空が占めているという特徴がある．

図 11A の再現に取り組んだ際のスクリーンショットを図 12，制作した画像を図 11B に示す．図 11B は山と空，手前に雑木林があり，画像の半分ほどを空が占めるという図

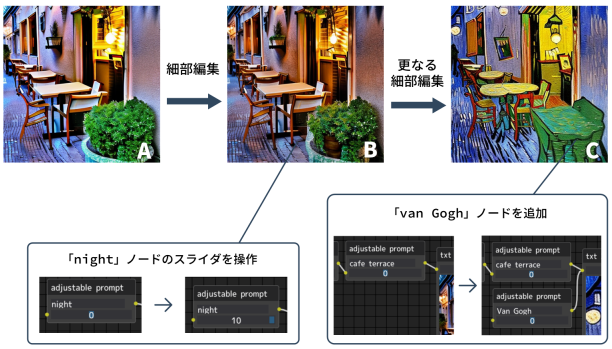

図 9 提案システムによって 1888 年のものと異なる「夜のカフェテラス」を制作した際に生成した画像．いずれの画像も図 8 内で生成されている画像とは大きく異なる構図を持つ．

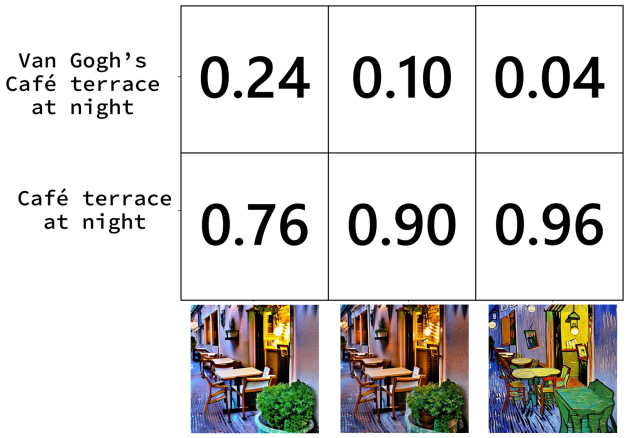

図 10 図 9A・B・C と 2 つのテキストを対象に算出した CLIP スコア．3 つの画像すべてが「van Gogh's cafe terrace at night」とのスコアが低いことがわかる．

11A と同じ特徴を持つ．しかし，提案システムによる細部編集にもかかわらず完全な再現はできていない．図 11B は図 11A と比較して山の標高や積雪の量と位置，雑木林の色などが異なっている．

細部編集をしたにもかかわらず元画像の再現ができなかった理由として，再現したい画像の構図を得ることの難しさがあると考えている．現状の提案システムでは，求めている構図を持つ画像が生成されるまでは生成を繰り返すしかない．構図を指定する内容のテキストを入力したノードを接続することである程度の構図を指定することは可能だが，とある画像と全く同じ構図をテキストのみで指定することは難しい．Stable Diffusion では画像を入力することが可能であり，入力された画像と同様の構図をもつ画像を生成することができる．

そこで，提案システムにも画像入力が可能となるよう，画像入力ノードを作成し画像生成を行った．その際の提案システムのスクリーンショットを図 13，制作した画像を図 11C に示す．なお，画像入力ノードは図 11A を再現するためだけに実装したため，他の画像を入力として設定することはできない．図 11B と図 11C を比較すると，画像

図 11 再現を試みた画像（A）と提案システムによって制作した再現画像（B・C）

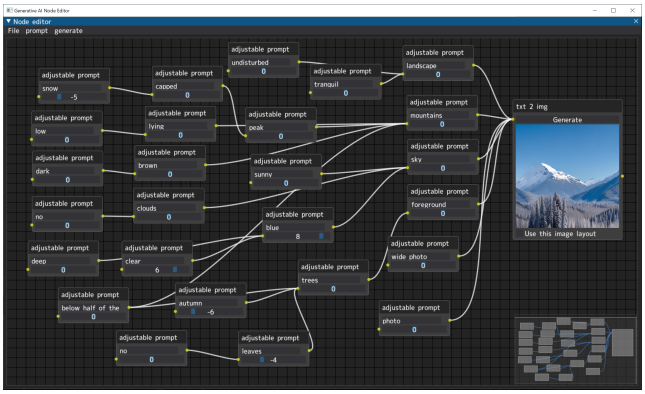

図 12 図 11B を制作した際の提案システムの様子.

入力により構図が図 11A に近づいていることに加え，雲の量が減り，オレンジ色の発色が弱まっていることが確認できた．実際に図 11C の方が図 11A に近づいたことを確認するため，SSIM（Structual Similarity）[13] を算出した．SSIM は画像の輝度・コントラスト・構造を考慮した画像の評価手法であり，同一の画像で算出すると 1 になる．図 11A をターゲットとし，図 11A・B・C との間で算出した SSIM を図 14 に示す．算出した値は小数第 6 位で切り捨てている．図 14 から，図 11C の方が 11B よりも高い値を示していることがわかる．だが，図 11C でも山や雑木林の全体的な質感や空の色などは再現できていない．そのため，提案システムによって完全に同じ画像を再現するまで追い込むことができるとはいえない．しかし，画像入力ノードを追加することでより類似した画像の制作に成功したことから，他の画像編集手法や Stable Diffusion の機能などをノードの追加によって取り入れ，システムを拡張することで完全な再現となるまで追い込めるようになる可能性があると著者は考えている．

5. 議論

提案システムはノードベースシステムであるため，画像の生成を行う際はノードの追加，テキストの入力とノードの接続という 3 つの操作を要する．だが，Stable Diffusion web UI [14] などの一般的な Stable Diffusion を扱うシステ

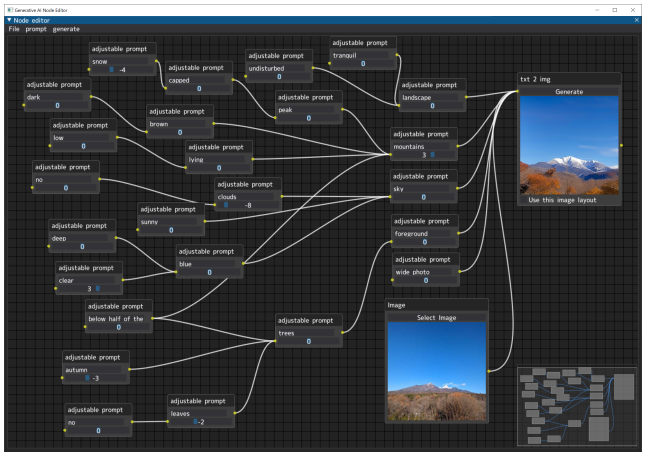

図 13 図 11C を制作した際の提案システムの様子.

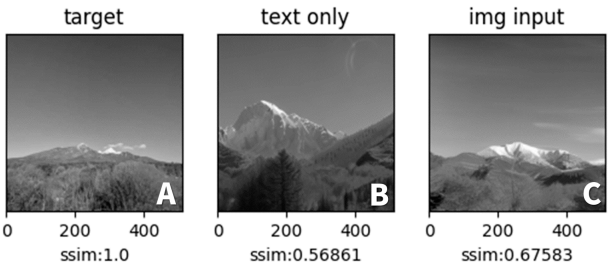

図 14 図 11A を対象画像として算出した SSIM．図 11B よりも図 11C の方が図 11A との類似度が高く，構造的に類似していることがわかる．

ムはテキストボックスへのテキスト入力という 1 つの操作だけで画像生成が行える．よって，画像生成だけを対象にすると提案システムの方が従来のシステムより必要な操作が多く，手間が増えている．だが，画像編集も対象にすると手間が減少する．既存の画像編集手法を利用するためには生成した画像のパスや使用したプロンプト，プロンプト内のどの単語に対してどのような操作を行うかをまとめたデータやプログラムの作成が必要となる．一方，提案システムでは画像編集も最大で 3 つの操作にとどまる．

画像生成時における手間を減少させるためには平文のプロンプトから自動でノードの構造を決定する仕組みの導入が必要であると考えている．具体的には提案システムに入力されたテキストを自動で解析し，適切なノード構造にしてシステムに反映させるノードを追加し，平文のプロンプトも入力可能にすることを想定している．平文でのプロンプト入力も可能にすることで画像生成にかかる手間も同程度になると考えられるが，現時点ではそのような仕組みは実装していない．

また，提案手法はノードの接続によってプロンプトを構築する方式のため，Stable Diffusion Prompt Book [15] などのユーザ間で共有されているプロンプトのノウハウを

そのまま使うことは困難である．画像の画風を変化させるための単語など，単語にフォーカスしたノウハウは提案システムでもそのまま利用可能だが，プロンプト例をそのまま使うことが難しい．よって，プロンプト例をそのままコピーしてペーストすればよい従来のシステムとは異なり，提案システムを使い始めるときにはユーザ自身がはじめからプロンプトを構築する必要がある．そのため，提案システムは使い始めのハードルが既存のシステムよりも高い．使い始めのハードルを下げるためには予め作成したテンプレートをインポートできる機能の実装などが必要であるが，現時点ではそのような機能はない．

6. 展望

本研究ではテキストによる画像生成においてユーザ自身が求める表現へと追い込めるようになることを目的とし，ノードやスライダの操作で画像の生成と編集が行えるシステムを提案した．更に，提案システムを用いて作例を制作し，CLIP スコアや SSIM などによる評価を行った．こうした数値による評価と作例制作を通して得た知見から，提案システムによってユーザが求める表現を追い込めるようになることを示した．

今後は様々な Generative AI を提案システムのようにノードベースのシステムで利用可能にすることで，様々な AI を統合した環境に発展させていきたいと考えている．現状，GPT-4 や Stable Diffusion といった Generative AI を利用するにはそれぞれのシステムを使い分けなければならない．だが，提案システムのようにそれぞれの AI をノードとして管理し，ユーザが自由に使い分けたり組み合わせたりできる環境が必要であると考える．本研究は複数の Generative AI を統合したシステムの第一歩として，GPT-4 と Stable Diffusion を同時に利用することによる画像生成と編集のためのシステムを提案した．生成するコンテンツごとに異なるモデルを個別のシステムを切り替えながら利用するのではなく，これらを単一のシステムで利用できるようにすることで人間と AI，AI と AI 同士の掛け合わせによる新たなコンテンツの生成が可能になることが期待できる．

参考文献

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis With Latent Diffusion Models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695 (2022).

[2] Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L. and Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, *ACM Transactions on Graphics (TOG)*, Vol. 42, No. 4, pp. 1–10 (2023).

[3] Ho, J., Jain, A. and Abbeel, P.: Denoising Diffusion Probabilistic Models, *Advances in Neural Information Processing Systems* (Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. and Lin, H., eds.), Vol. 33, Curran Associates, Inc., available from ⟨`https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf`⟩ (2020).

[4] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y. and Cohen-or, D.: Prompt-to-Prompt Image Editing with Cross-Attention Control, *The Eleventh International Conference on Learning Representations*, available from ⟨https://openreview.net/forum?id=$_C Dixzkzeyb$⟩ (2023).

[5] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J. and Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, *Advances in Neural Information Processing Systems* (Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K. and Oh, A., eds.), Vol. 35, Curran Associates, Inc., available from ⟨`https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf`⟩ (2022).

[6] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M.: Hierarchical text-conditional image generation with clip latents.

[7] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N. and Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics, *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, JMLR.org, p. 2256–2265 (2015).

[8] comfyanonymous: ComfyUI, `https://github.com/comfyanonymous/ComfyUI`. (Accessed on 2024/2/23).

[9] Tumanyan, N., Geyer, M., Bagon, S. and Dekel, T.: Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930 (2023).

[10] Ge, S., Park, T., Zhu, J.-Y. and Huang, J.-B.: Expressive Text-to-Image Generation with Rich Text, *IEEE International Conference on Computer Vision (ICCV)* (2023).

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning transferable visual models from natural language supervision, *International conference on machine learning*, PMLR, pp. 8748–8763 (2021).

[12] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. et al.: Laion-5b: An open large-scale dataset for training next generation image-text models, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 25278–25294 (2022).

[13] Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P.: Image quality assessment: from error visibility to structural similarity, *IEEE transactions on image processing*, Vol. 13, No. 4, pp. 600–612 (2004).

[14] AUTOMATIC1111: Stable Diffusion web UI, `https://github.com/AUTOMATIC1111/stable-diffusion-webui`. (Accessed on 2024/2/23).

[15] OpenArt: Stable Diffusion Prompt Book, `https://openart.ai/promptbook`. (Accessed on 2024/2/23).