ホッピングマスクを利用した Stable Diffusion Model による連続画像生成

趙 開顔 [1]　Qiyu Wu[1]　村田 秀樹 [2]
石川 隆一 [2]　根之木 颯亮 [2] 山本 覚 [2]　鶴岡 慶雅 [1]
[1] 東京大学大学院 [2] 株式会社電通デジタル
{zhaokaiyan1006,qiyuw,yoshimasa-tsuruoka}@g.ecc.u-tokyo.ac.jp
{murata.h,ishikawa.r,nenoki.s,yamamoto.sato}@dentsudigital.co.jp

概要

Stable Diffusion Model (SDM) のオープンソース化に伴い，大規模画像生成モデル，特に text-to-image モデルが話題になっている．既存研究では拡散モデルを改善し，制御可能な画像生成や画像編集を目指している．これらの方法はいずれも有効だが，使われるプロンプトに対する制約が大きく，編集した画像の構成を変えられないなどの欠点が存在する．それに加え，従来の SDM は，生成した画像の表現力が足りず，しばしば品質の低い画像が出力されるなどの問題点を抱えている．これらを踏まえて，本研究では再学習を必要とせず，連続的に生成される画像間の相関性を改善し，画質や表現力を向上させる手法を提案する[1)]．

1　はじめに

近年，拡散モデル（DDPM）[1] の幅広い応用や CLIP [2] などのクロスモーダルモデルの発展に伴い，Stable Diffusion Model [3]，DALLE2 [4]，Imagen [5] などの大規模画像生成モデルが話題になっている．これらのモデルにより，専門知識の有無を問わず，誰でも高品質な画像を生成することができるようになった．図 1 に示すように，プロンプト（生成する画像の内容を示すテキスト）さえあれば，モデルが自動的にプロンプトが表す内容を画像で出力することが可能である．このようなテキストから画像を生成する技術は txt2img(text-to-image) と呼ばれている．

2022 年の 8 月，SDM のオープンソース化を機に，画像生成への関心はさらに高まった．画像生成を改善する方向の一つとして，制御可能性を高める方法

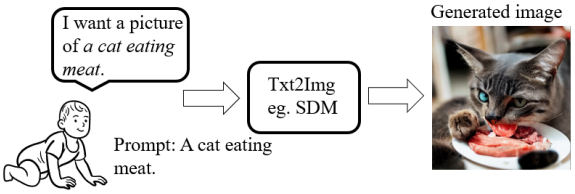

図 1　txt2img の例

が挙げられる．Hertz ら [6] の prompt-to-prompt という手法により，プロンプトのテキスト編集だけで画像を編集することが可能になった．一方，Kawar ら [7] は一枚の画像と一つのプロンプトだけで，複雑な画像編集を可能にする手法を提案した．Stable Diffusion Model (SDM) への関心を踏まえ，本研究においても SDM である txt2img モデルを使用した．

しかし，Yang ら [8] によると，従来の SDM にはまだ多くの問題が残っている．まず，SDM は高品質な画像を生成することができるものの，時には学習がうまくいかず，崩れた画像を生成してしまうことがよくある [9]．また，図 2 に示したように，入力したプロンプトは「ジョンがバットを振っている」であるのに，出力された画像は「バットを振っている」という動きをうまく把握していない場合が存在する．これは，生成画像がプロンプトに忠実でないという問題である．そのほかにも，複数の入力（幾つかのプロンプトを同時に入力する場合）に対しては，プロンプト数を認識する機能がなく，叙事的な画像生成 (narrative image generation) は不可能である．これらの問題を踏まえ，私たちはホッピングマスクを提案し，CLIP のエンコーダー側で上記の問題点を改善した．

Wu ら [10] と Huang ら [11] は文間の情報 (cross-sentence information) を利用し，入力文のセマンティック情報を強化することで，言語モデルとビデオローカライゼーションの精度を向上させた．こう

1) コードは https://github.com/KYuuto1006/sdm_hopping_mask で公開されている．

図 2　SDM が生成した画像がプロンプトに忠実でない例

した手法から着想を得て、我々の提案手法においても文間の情報を利用することで，高品質，高表現力の画像連続生成を実現した．

実験では，同じプロンプト，シードで提案手法により生成した画像と従来の SDM で生成された画像を比較したところ，文節間情報の重要性が示された．

本研究での貢献は以下の通りである：

- 分離符 $< SEP >$ の使用により，複数のプロンプトを同時に SDM に入力しても，プロンプト数を認識することが可能になった．
- ホッピングマスクを提案し，文間の情報を使用することにより，画像の表現力が向上．
- 提案したホッピングマスクにより，一定の一貫性がある画像を生成することができた．

2　関連研究

2.1　txt2img 画像生成

テキストから画像の生成という分野は，2016 年の Reed ら [12] の研究から始まった．テキスト記述に基づいた画像を生成するために従来の GAN [13] を拡張し，小規模なデータセットと小さな画像解像度で動作することが示された．以来数年間，主な txt2img モデルは GAN のネットワーク構造 (Generator と Discriminator 二つのネットワークを互いに競争させるような形で学習する) に従い，生成画像の品質と解像度の両方を向上させたが，近年話題になっているものは拡散モデルである．

拡散モデル [14] が提唱されたのは数年前に遡ることができるが，2020 年，Ho ら [1] が初めて何故拡散モデルは GAN よりも性能が良いのかに関し数値的な分析を行った．拡散モデルの仕組みは非常にシンプルで，まず画像にノイズを付与し，テキストなどの他のモダリティからの情報を通し，ノイズを除去するものである．

しかし，画像をピクセルの配列として直接扱う場合，変数の次元が多くなり，計算量が増加してしまうという問題が存在する．ここで，SDM [3] が提案された．SDM は，ノイズから画像を生成する代わりに，ノイズから潜在表現 (latent) をまず生成し，それを画像へと戻すという 2 段階のプロセスを経て画像を生成するモデルである．

2.2　拡散モデルによる画像編集

従来の SDM はすでに高品質な画像を生成することができるが，画像生成を改善する方向の一つとして，制御可能な画像生成と画像編集が挙げられる．

Hertz ら [6] からの prompt-to-prompt 手法により，プロンプトのテキスト編集だけで画像を編集することが可能になった．Prompt-to-prompt のキーポイントは，ピクセルとトークンの関係を利用し，クロスアテンションマップを拡散過程に埋め込み，生成を制御することである．他の手法と異なり，学習，微調整，追加データや最適化を必要とせず，単に入力プロンプトを変更するだけで，画像の背景を変更せずに小さな部分を変更することや，画像のスタイルを全体に変更することが可能になった．また，新しい情報を追加することも可能になった．しかし，この方法での入力プロンプトに対する制約は大きく，プロンプト間のギャップは数語または数フレーズでなければならない．

一方，Kawar ら [7] は Imagic を提案した．Imagic は，初めて 1 枚の実画像を，テキストのみを用いて複雑な編集をすることを可能にした．鮮明な画像さえあれば，Imagic は犬を座らせ，又は，ジャンプさせる，また，鳥を羽ばたかせることができる．だが，この方法による編集にも限界があり，編集した画像の構成を変えることはできない．

これらの方法はいずれも有効だが，画像の表現力が不十分な点や，画像編集が制約される等の問題は解決されていない．本研究において，私たちは学習を必要とせず，連続的に生成される画像間の相関性を改善し，画質や表現力を向上させる手法を提案する．

3　提案手法

3.1　SDM による画像生成

提案手法を紹介する前に，従来の SDM によるプロンプトから画像への変換の流れについて説明する．

図 3 左側のように，一つのプロンプトを，まず CLIP [2] のトークナイザー (tokenizer) に入力し，トークン，$T_1, T_2, T_3, ...$，への変換を行う．次に，式 (1) の通り，CLIP エンコーダーを使用し，トークンを埋め込みに変換する．

$$e_i = Attn(w_i) \tag{1}$$

ここで，w_i は i 個目のトークンの単語埋め込み (word embedding) を示し，e_i がアテンションの計算を行ったトークンの埋め込みである．この変換の中で、コーザルマスク (causal mask) [15] が SDM に使用されている．図 4 の左側に示すように，第一列と第一行は各トークンで，第一列から見ると，青い部分はアテンションを計算する際，計算に含まれることを示す．例えば，T_1 は，自分のみを考慮し，T_2 は，T_1 と自身に基づき，アテンションを計算する．つまり，トークンがアテンションを計算する際，その前のトークンと自身のみを認識できるようにしている．このような仕組みを使用した原因としては，トークンの後に，意味のないパディング (padding) がついているので，意味のあるトークンをパディングを見えないように設定にしている．

続いて，埋め込み e をノイズを加えられた潜在表現 (latent) と一緒に U-Net[16] に入力し，埋め込みを基にノイズを除去する:

$$\hat{l} = UNet(l, e) \tag{2}$$

l はノイズを加えられた潜在表現であり，$\hat{l}$ はノイズを除去した潜在表現である．最後に，ノイズが除去された潜在表現を用い，VAE [17] を通し，ベクトルから画像へ戻す．以下の式で表すことができる．

$$image = vae(\hat{l}) \tag{3}$$

3.2　ホッピングマスクを用いた SDM

文間の情報を利用するために，まず，入力をいくつかの連続プロンプトにする必要がある．モデルが一連の文を明示的にエンコードできるように、セパレータ $< SEP >$ を文に埋め込む．この際，提案した

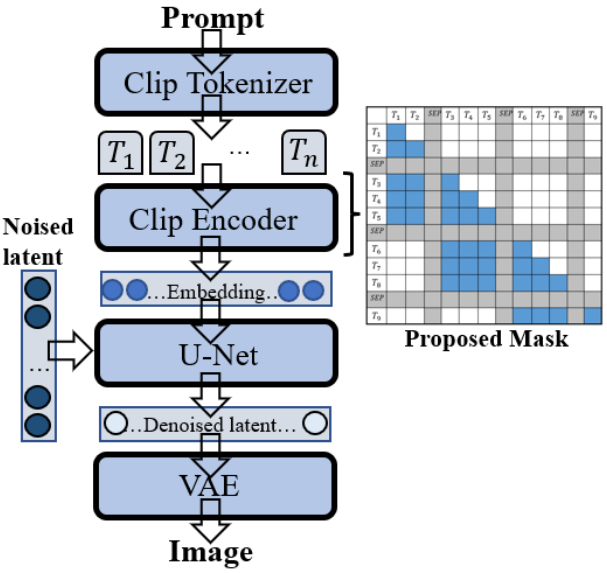

図 3　SDM 及び提案手法の流れ

手法の入力は以下の式で表すことができる：

$$\begin{aligned} x &= Prompt_1 < SEP > Prompt_2 < SEP > ..., \\ Prompt_1 &= T_1, ..., T_n \end{aligned} \tag{4}$$

トークン化後，複数のプロンプトが存在するため，トークンが所属するプロンプトの ID を返す関数 $P_{id}(\cdot)$ を使う．例えば，T_8 が第三個目のプロンプト ($Prompt_3$) に属する場合，$P_{id}(8)$ の値は 3 となる．

続いて，我々が提案するホッピングマスクについて説明する．$Prompt_2$ の埋め込み (embedding) を計算する際，$Prompt_1$ の情報を踏まえて，アテンションを計算していく．そして，$Prompt_3$ の埋め込みを計算する際，今回の場合 $Prompt_2$ の情報を考慮し，アテンションの計算を行う．図 4 の右側に示す通り，例えば，$Prompt_3$ に所属している T_8 を考えると，T_8 のアテンションの計算には，$Prompt_2$ のトークンと $Prompt_3$ の中で，T_9 自分と自分より前のトークンを考慮し，$Prompt_1$ のトークンは含まれていない．定式化すると，以下のように表される：

$$M_{i,j} = \begin{cases} 1 & i \le j, P_{id}(i) - P_{id}(j) \le 1 \\ 0 & \text{ほか} \end{cases} \tag{5}$$

$M_{i,j}$ は，提案したアテンション行列の i 行目，j 列目のマスク値のことであり，青い部分の値は 1 である．分離符として使われた $< SEP >$ も計算時に考慮する必要がないため，ゼロに設定している．

次に，式 (1) に従い，CLIP エンコーダーでのアテンションの計算結果は以下のようになる：

$$Attn = Attn(w_i) * M \tag{6}$$

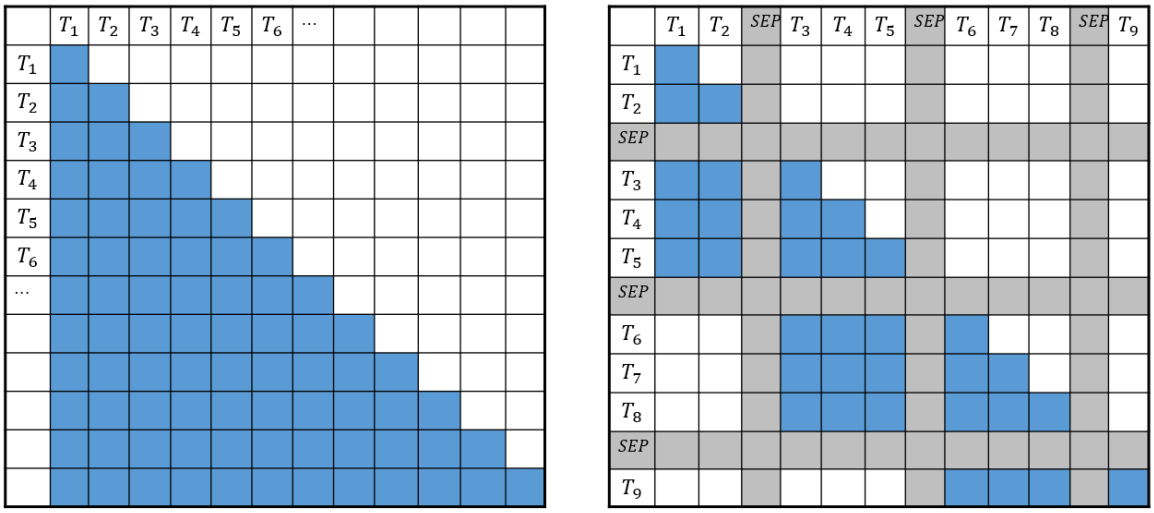

図 4　左側は SDM に使われているコーザルマスクで，右側は提案したホッピングマスクである．

図 5　SDM が生成した画像と提案手法による生成画像の比較

このようなアテンション計算がエンコーダーの各層で行われ，最終的に得られるのは複数のプロンプトからホッピングマスクを使用した埋め込みである．SDM の手順に従い，次に U-Net を使用し，ノイズを除去する際には，入力の各自プロンプトの埋め込みを抽出し，入力プロンプト数と対応している数の潜在表現から，別々で画像を生成する：

$$\hat{l}_i = vae(l, e_i), i \in P_{id}(\cdot) \tag{7}$$

提案手法により，豊富な文脈情報の取得に成功し，SDM に基づいた生成画像の品質と表現力の向上だけでなく，連続画像生成が可能になった．

3.3　分析

Wu ら [10] と Huang ら [11] によると，文間の情報は，モデルがトークンとプロンプトの意味を正しく理解することに非常に役に立っているとされる．私たちもホッピングマスクを使用することで，文間の情報を利用し，プロンプトの埋め込みを改善した．

図 5 の右の画像は，「John throws out a baseball」の後に生成された 2 番目の画像である．右の画像は左の画像と比較し，遥かに表現力があることがわかる．これは，2 つ目のプロンプトの埋め込みを計算する際に，1 つ目のプロンプトの情報も考慮することで，より入力プロンプトに適合した画像を生成する可能性を高めているためである．つまり，「John throws out a baseball」と「John is waving a baseball bat」は，関連する或いは同じ活動を記述しているため，画像 2 を生成する際に，ジョンが野球をしているという事実が描写される保証が高くなると考えられる．

さらに，提案手法を用いることにより，連続的に生成される画像間の一貫性も向上にも繋がる．これは，異なるプロンプトで同じ名前の人物が現れる場合，後者の名前のトークンは前者のものを参照ことができるため，類似した人物が生成されやすくなったと解釈できる．その他実験結果については付録を参照していただきたい．

4　まとめ

本研究では再学習を必要とせず，連続的に生成される画像間の相関性を改善し，画質や表現力を向上させる手法を提案した．既存の画像編集方法と比較し，プロンプトの制約が減少し，連続生成した画像の一貫性の向上に成功した．

謝辞

この研究は東京大学 AI センターと株式会社電通デジタルが推進する共同研究の助成を受けています.

参考文献

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 6840–6851. Curran Associates, Inc., 2020.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International Conference on Machine Learning**, pp. 8748–8763. PMLR, 2021.

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

[4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. **arXiv preprint arXiv:2204.06125**, 2022.

[5] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding, 2022. **URL https://arxiv. org/abs/2205.11487**, Vol. 4, .

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. **arXiv preprint arXiv:2208.01626**, 2022.

[7] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. **arXiv preprint arXiv:2210.09276**, 2022.

[8] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. **arXiv preprint arXiv:2209.00796**, 2022.

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. **arXiv preprint arXiv:2208.01618**, 2022.

[10] Qiyu Wu, Chen Xing, Yatao Li, Guolin Ke, Di He, and Tie-Yan Liu. Taking notes on the fly helps language pre-training. In **International Conference on Learning Representations**, 2020.

[11] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 7199–7208, 2021.

[12] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In **International conference on machine learning**, pp. 1060–1069. PMLR, 2016.

[13] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. **arXiv preprint arXiv:1411.1784**, 2014.

[14] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In **International Conference on Machine Learning**, pp. 2256–2265. PMLR, 2015.

[15] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 9847–9857, 2021.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In **International Conference on Medical image computing and computer-assisted intervention**, pp. 234–241. Springer, 2015.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. **arXiv preprint arXiv:1312.6114**, 2013.

A　付録 (Appendix)

A.1　アテンションマスクの可視化

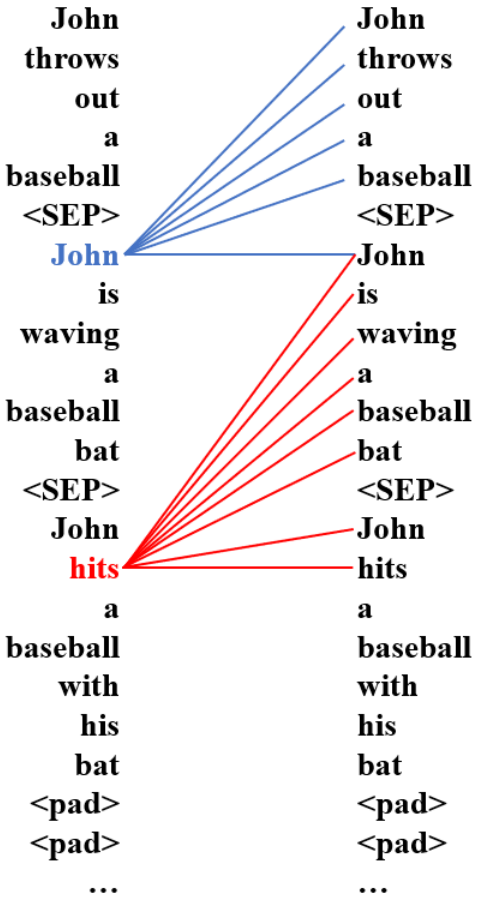

図 6　提案マスクの可視化表示

A.2　実験設定

今回使用した SDM モデルは，stable-diffsuion-v1-4[2)]である．生成された画像の解像度は 512*512 で，U-Net のステップ数は 50 に設定されている．

A.3　他の実験結果

実験結果に関しては，比較を行った対象は同じシードを使った生成結果である．

A.3.1　表現力向上の例

図 7 の中で，左側のほうは従来の SDM が生成した画像であり，右側のは提案手法の生成結果である．上から一組目は John is working in his farm. の次に生成された二番目の画像だ．二組目は John throws out a baseball.< SEP >John is waving a baseball bat.< SEP >John hits a ball with his bat. の次の結果である．三組目は A mountain is tall and steep. の次に生成された二番目の画像.

図 7　表現力向上の例

A.3.2　一貫性向上の例

図 8　一貫性向上の例

図 8 の上列は提案手法により生成した画像であり，犬の様子は下の SDM が生成した画像より同じふうに見える．

2)　https://huggingface.co/CompVis/stable-diffusion-v1-4