潜在拡散モデルにおける生成対象の個別配置と生成による画質改善の試み

Feasibility Study on Improving Image Quality by Individually Placing and Generating Objects in Latent Diffusion Models

永井大地 [1] 守田竜梧 [2] 北田俊輔 [2] 彌冨 仁 [1,2]

法政大学理工学部応用情報工学科 [1]　法政大学理工学研究科応用情報工学専攻 [2]

概 要

潜在拡散モデル（LDM）は、プロンプトから高品質な画像を生成するが、複数物体を含むプロンプトについては物体の欠落や属性の混同が課題となる。本研究では、逆拡散過程で物体の存在が確定するステップに着目した新たな枠組 TAUE(Training-free trAnsplant cUltivation diffusion modEl) を提案する。TAUE では各物体と背景を個別に生成し、苗を畑に植えるように、物体の存在が確定したノイズで背景上の対象領域を上書きすることで物体の欠落を防ぎ、配置に忠実な画像の生成を実現する。TAUE は様々な LDM の関連技術への応用が期待され、直感的な生成プロセスを提供する。

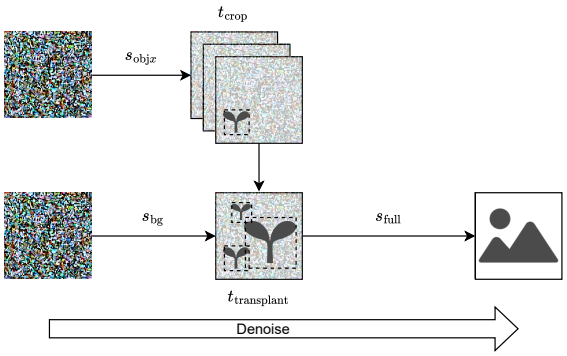

図 1: TAUE の概要図

1　背景

近年の画像生成モデルの進歩と普及は目覚ましく、その生成画像は文化的な活用から科学の発展まで、様々なフィールドで活躍している。代表的なモデルである Latent Diffusion Model (LDM)[1] は初期ノイズ（ガウスノイズ）から、逆拡散過程と呼ばれるノイズ除去の処理を経て、画像を生成する。Stability AI 社の Stable Diffusion (SD) は 、データセット LAION-5B [2] による約 50 億の画像とテキストのペアを学習した LDM であり、高品質な画像を生成可能である。一方で、複数の生成対象を持つプロンプトに対し、しばしばオブジェクトの欠落や対象が持つ属性同士の混同が発生する。本事象に対する最も有効な改善策は、学習データの強化とそれを用いた再学習が挙げられるが、高コストでありトレーニングフリーの改善手法が求められる。

LDM による画像生成において発生する、生成対象の欠落や属性の割り当てミスを改善するため、多くの研究がされてきた。StructuredDiffusion[3] は、名詞に係る属性情報の cross attention [4] を工夫するという言語的なアプローチにより、属性の割り当てミスに対する改善を試みた。しかし、手法適用前後で生成される画像が類似しており、結果として改善効果が低下するという問題がある。また、Attend & Exciten [5] は生成対象の欠落に対して、生成対象を表すトークンに対する cross attention が、少なくとも 1 つのパッチに割り当てられることを保証する機構を導入する方法を示した。生成対象の欠落に一定の改善を示しつつも非現実的な組み合わせに対しリアリティを失う等の問題が残り、完全な解決には至っていない。

そこで、我々は逆拡散過程において生成対象の存在が決定するタイミングに着目した新たな枠組 TAUE (Training-free trAnsplant cUltivation diffusion modEl) を提案する。本枠組では、それぞれの生成対象はプロンプトをもとに、異なる潜在表現として中間段階まで生成される。こうして生成された中間段階の潜在表現は、任意の潜在表現に移植することで生成時に意図された物体を発現させることができ、植物の苗のように働くことから「苗ノイズ」と呼称する。本枠組は LDM に関連する多くの技術への応用が期待され、本稿ではその 1 つとして Layout to Image（L2I）タスクへの応用を行った。実験セクションでは生成対象を 2 つ以上含むプロンプトを用いて生成実験を行い、生成対象の欠落防止を試みた。

2　TAUE

提案手法の概要図を図 1 に示す。我々は物体の発現能力を持つ苗ノイズを構築したのち、背景として生成した潜在表現に移植する 2 段階の画像生成パイプラインを提案する。

2.1　苗ノイズの構築

苗ノイズ z_{seedling} は生成対象を発現させるノイズとして生成され、L2I の要求を満たすために位置やサイズに関する条件を受けながら生成される。これは生成対象の Green Back(GB) 画像を生成する中間段階 t_{crop} でのノイズとして得られる。なお、この手法は Morita らの Diffusion モデルを用いたクロマキー生成手法 [6] に着想を得ており、サイズや位置に柔軟に対応するよう GB 成分の注入方法を調整している。

2.2　苗ノイズの移植

移植では、背景情報のプロンプト s_{bg} で中間段階 $t_{\text{transplant}}$ まで生成された潜在表現 z_{bg} の対象領域を、

取得した苗ノイズで上書きすることで適切かつ直感的な物体配置を行う。移植後はプロンプトをすべての生成対象と背景情報を含むプロンプト s_{full} に変更し、逆拡散過程を継続する。実験セクションでは $t_{\text{transplant}} = 15$ の場合と $t_{\text{transplant}} = 0$ の場合について実験を行った。ただし、後者の場合では以下の式の通り、苗ノイズにスケーリングを施している。

$$z'_{\text{seedling}} = \frac{\sigma_{z_{\text{bg}}}}{\sigma_{z_{\text{seedling}}}} \cdot z_{\text{seedling}}$$

2.3 Attention Assignment Score(AAS)

既存の評価指標はプロンプト等の生成条件に対する生成画像の評価が主である。一方で拡散モデルは逐次的に画像が構築されるため、その過程における挙動は重要な評価対象だと考える。そこで我々が提案する AAS は生成過程における生成対象に対する cross attention と、それに付随する属性単語に対する cross attention の一致度を測る指標である。これは attention の割り当ての成否と生成画像における属性割り当ての成否という 2 つの状況のギャップを数値化し、画像生成での属性割り当てミス等の問題を細分化する。（設計中でイメージが固まっていません）

$$AAS = \sum_{obj \in objs} \text{定義式}$$

エンコーダは式 1 入力 x を低次元の潜在表現 $z(x)$ に圧縮する。

$$z(x) = f(x; W) \tag{1}$$

一方、デコーダは式 2 により潜在表現 z(x) を元の空間に $x'(z)$ として復元する。

$$x(z) = \hat{f}(z; \hat{W}) \tag{2}$$

1. **生成に適した領域の探索**はじめに生成物体である“cat”の attention を可視化し、後述するフィルターを畳み込み、最大値をとる index を求める。この操作は atttention のパターンがフィルタのパターンに最も一致する領域を探索することと解釈できる。この操作は式 **??**で表される。

$$(i, j) = \text{argmax}_{(i,j)}(\sum_{m,n} \text{Attention}(i+m, j+n) \cdot \text{kernel}(m, n)) \tag{3}$$

フィルターは物体の生成領域を指定する BB と同じ形状を持ち、値は中央からの距離に応じて減少し、境界付近では負の値になる。このフィルタの設計意図は選択領域の中央が attention が高く、逆に境界周辺では低い様な領域を選択するようになっている。これにより、BB から生成物体がはみ出すことが抑制される。

2.

表 1: 結果

	IOU	CLIP Score	AAS
Ours($t_{transplant} = 0$)	0.000	0.000	0.000
Ours($t_{transplant} = 15$)	0.000	0.000	0.000
Some Pipeline	0.000	0.000	0.000

3 実験

本研究では、LDM ベースのモデルとして Stable Diffusion XL (SDXL) [1] を使用した。プロンプトには、DrawBench データセット [7] の中から Colors カテゴリに属し、複数の生成対象を含むものを選定した。比較実験では、提案手法（$t_{\text{transplant}} = 0, 15$）および XXX（標準的な L2I パイプライン）の 3 つの条件下で、各プロンプトに対して 100 枚ずつ画像を生成した。生成結果の評価には、以下の 3 つの指標を使用した：

- Intersection over Union (IoU)
- CLIP Score [8]
- Attention Assignment Score (AAS)

4 結果と考察

結果を表 1 に示す

5 結論と今後の展望

参考文献

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *In Proc. of CVPR*, pp. 10684–10695, 2022.

[2] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *In Proc. of NeurIPS*, Vol. 35, pp. 25278–25294, 2022.

[3] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *In Proc. of ICLR*, 2022.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *In Proc. of NeurIPS*, p. 6000–6010, 2017.

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *In Proc. of SIGGRAPH*, Vol. 42, No. 4, pp. 1–10, 2023.

[6] Ryugo Morita, Stanislav Frolov, Brian Bernhard Moser, Takahiro Shirakawa, Ko Watanabe, Andreas Dengel, and Jinjia Zhou. Tkg-dm: Training-free chroma key content generation diffusion model. *arXiv preprint arXiv:2411.15580*, 2024.

[7] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *In Proc. of NeurIPS*, Vol. 35, pp. 36479–36494, 2022.

[8] 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 *In Proc. of ICML*, pp. 8748–8763, 2021.

[1] https://hf.co/stabilityai/stable-diffusion-xl-base-1.0