

立教大学学術推進特別重点資金（立教 S F R）

大学院学生研究

2 0 2 5 年度研究成果報告書

研究科名	立教大学大学院 人工知能科学研究科 人工知能科学専攻		
研究代表者 (2026 年 3 月現在 のものを記入)	在籍課程・学年	氏 名	
	☐ 博士前期課程 年 ☒ 博士後期課程 1 年	倉 員 豪	
指導教員	所属部局・職名	氏 名	
	人工知能科学研究科・ 特任准教授	石 川 真之介	
自然・人文 ・社会の別	☐ 自然 ・ 人文 ・ 社会	個人・共同の別	☐ 個人 ・ 共同 名
研究課題	Vision & Language モデルを用いた画像生成における視覚的表現の最適化		
研究組織 (共同研究者が いる場合のみ、 2026 年 3 月現在 のものを記入)	在籍研究科・専攻・課程・学年	氏 名	
研究期間	2 0 2 5 年度		
研究経費 (1 円単位)	(支出金額) 190,800 円 / (採択金額) 200,000 円		

研究の概要 (200~300 字で記入、図・グラフ等は使用しないこと。)
当該研究の研究目的を含むこと。 テキストから画像を生成する拡散モデルは高品質な画像を生成できるが、「左のマグカップは赤、右のマグカップは青」のように、指定した属性を正しい位置の物体に割り当てる空間的属性バインディングにおいてしばしば失敗する。本研究では、当初計画していた初期ノイズ画像の編集アプローチを出発点として探索を行う中で、cross-attention の logit に空間バイアスを直接加算する training-free な推論時手法 SLM(Spatial Logit Modulation)を開発した。画像生成の失敗を Mixing と Swap に分解し、5 カテゴリ 500 枚の独自ベンチマークにおいて、Swap を 59%削減し Precision を 48.1%から 72.6%に改善した。本成果は国際会議向け英文論文として取りまとめ、現在査読中である。
キーワード (研究内容をよく表しているものを3項目以内で記入。)
[深層学習] [生成 AI] [画像生成]

研究成果の概要 (図・グラフ等は使用しないこと。)

以下の視点を含めて記載のこと。

- ・当該研究は何をどこまで明らかにできたのか (できなかったのか)。
- ・何をもって研究成果 (経過) を達成できた (できなかった) と考えられるのか。
自身が設定した研究目的・目標に照らして、その根拠がわかるよう記載のこと。

(1) 研究目的

本研究は、申請時に掲げた「Vision & Language モデルを用いた画像生成における視覚的表現の最適化」を目的とし、テキストプロンプトに基づく画像生成において、ユーザーの意図する視覚的表現を忠実に再現する手法の開発を目指した。当初は初期ノイズ画像の編集による構図・スタイル制御を計画していたが、先行研究である InitN0 の再現実験および拡散モデル内部の cross-attention 構造の分析を通じて、空間的属性バインディング (指定した色を正しい位置の物体に割り当てる) という具体的な課題に焦点を絞った。

(2) 研究の経緯と方向転換

フェーズ 1 (4~9 月) では、申請時の計画に沿い、InitN0 の再現実装を行うとともに、Crystal Ball などの先行研究を参考に基礎検証を実施した。InitN0 の cross-attention スコアに基づくノイズ最適化を検証する過程で、Attend & Excite や Prompt to Prompt などにより cross-attention の attention map が物体の空間配置に強く寄与していることを確認した。また、画像編集をする際に、Cross-Attention だけではなく、Self-Attention も併用した研究が見られることがわかった。

これら知見を踏まえ、フェーズ 2 (10~12 月) では、先行研究 SCDG (Self-Cross Diffusion Guidance for Text-to-Image Synthesis of Similar Subjects) をベースとした SCDG++ を構築し、self-attention と cross-attention の重なりを損失として設計する手法を検討した。この過程で、属性バインディングの失敗が「混色・漏れ (Mixing)」と「左右逆転 (Swap)」という性質の異なる二種類のモードに分解できることを実験的に確認した。SCDG++ は Mixing の低減には有効であったが、Swap の解消には至らなかった。

Swap の原因をさらに掘り下げるため、フェーズ 3 (1 月~) では、ISAC (ISAC: Training-Free Instance-to-Semantic Attention Control for Improving Multi-Instance Generation) 等の先行研究を参考に self-attention ベースのインスタンス領域推定と、領域と色トークンの対応分析を導入した。CLIP のテキスト・画像埋め込み解析と合わせて、Swap の一因となり得る CLIP-based conditioning の左右分離の弱さを分析するとともに、attention map に押すべき空間バイアスの信号が本来弱いことを確認した。この知見から、self/cross attention 損失の最適化ではなく cross-attention の logit に直接空間バイアスを加算するという SLM の着想に至った。この方向転換は、申請時に記載した「研究計画が想定通り進まなかった場合は柔軟にアプローチを修正する」方針に基づくものである。

(3) 失敗モードの分析

属性バインディングの失敗を、色の分離が不十分な Mixing と、色は分離されるが左右位置が反転する Swap の二つに分解する枠組みを提案した。既存手法である SCDG や CONFORM は self-cross guidance や contrastive loss により色の分離を促進するが、Swap の解消には直接寄与しないことを示した。

研究成果の概要 (つづき)

(4) 提案手法: Spatial Logit Modulation (SLM)

上記の分析に基づき、cross-attention の logit に空間バイアスを直接加算する手法 SLM を提案した。SLM は画像の左半分の領域が左側の属性トークンに注意を向け、右半分が右側の属性トークンに注意を向けるように、attention logit を増強する。この操作は training-free であり、推論時に追加の学習を必要としない。さらに、SLM と InitN0 に着想を得た iterative refinement を組み合わせる SLM+Refine を構成し、separation の補完を行った。

(5) 評価

5 カテゴリ (mug, car, bird, shoe, cat) × 2 つの色順序 × 50 seeds からなる合計 500 枚のベンチマーク (PromptBank) を構築し、Grounding DINO と CLIP に基づく自動評価パイプラインで定量評価を行った。SLM+Refine はベースラインの SD 1.5 に対して Swap を 66 件から 27 件に 59%削減し、Precision を 48.1%から 72.6%に改善した。また、既存手法との比較において、CONFORM がカテゴリによって性能が大きく変動する (bird で高性能だが car や shoe では不安定) のに対し、SLM+Refine はカテゴリ横断でより安定した改善を示した。

(6) 追加分析

補足的な分析として、(i) red/blue 以外の色ペアでも同じ方向の改善傾向が見られることを確認し、(ii) top/bottom 配置への拡張を検討した。また、バイアス強度 β 、refinement 強度 (scale_factor)、時間方向スケジューリングの感度分析を行い、設計選択の妥当性を検証した。特に scale_factor については、局所的なベンチマーク (Two Mugs) では強い設定 (sf=20) が最良であるが、多カテゴリ設定では保守的な設定 (sf=10) がロバストであることを確認した。

(7) 残された課題

cat カテゴリは全手法で困難なままであり、自然な毛色パターンと人工的な色指定の不一致が一因である可能性がある。また、本研究は Stable Diffusion 1.5 のみでの検証であり、SDXL や SD3 への拡張は今後の課題である。他にも、画像上の物体が 2 つの場合のみのケースを想定しており、3 つ以上のケースにも適用することを今後考えて行きたい。

研究発表 (研究によって得られた研究成果を発表した①～④について、該当するものを記入してください。該当するものが多い場合は主要なものを抜粋してください。なお、成果発表を確認できる資料を合わせて研究成果報告書提出フォームより提出してください (紙媒体等、研究成果報告書提出フォームから提出できない場合は、別途リサーチ・イニシアティブセンターへ提出してください)。

①雑誌論文 (著者名、論文標題、雑誌名、巻号、発行年、ページ)

②図書 (著者名、出版社、書名、発行年、総ページ数)

③シンポジウム・公開講演会等の開催 (会名、開催日、開催場所)

④その他 (学会発表、研究報告書の印刷等)

※修士論文・博士論文は含みません。

④ その他 (学会発表、研究報告書の印刷等)

倉員 豪、石川 真之介「Spatial Logit Modulation for Attribute Binding in Text-to-Image Diffusion Models」投稿先：ECCV 2026 (査読中)