MultiDiffusion: Diffusion Pathsの融合による制御可能な画像生成

Omer Bar-Tal、Lior Yariv、Yaron Lipman、Tali Dekel

概要

テキストから画像を生成する拡散モデルにおける最近の進歩により、画像品質の飛躍的な向上が実現しました。しかし、生成される画像に対するユーザーの制御性や、新しいタスクへの迅速な適応は依然として未解決の課題です。これらの課題は現在、高コストで時間のかかる再トレーニングや微調整、または特定の画像生成タスクへのアドホックな適応によって主に対処されています。本研究では、さらなるトレーニングや微調整を行うことなく、事前学習されたテキストから画像生成拡散モデルを用いて、多様で制御可能な画像生成を可能にする統一フレームワークである **MultiDiffusion** を提案します。

本アプローチの中心には、新しい生成プロセスがあります。これは、複数の拡散生成プロセスを共有パラメータや制約のもとで結合する最適化タスクに基づいています。MultiDiffusion は、ユーザーが指定した制御条件（例: 希望するアスペクト比や空間的な誘導信号）に従い、高品質で多様な画像を生成するために容易に適用可能であることを示します。この制御条件には、パノラマのような希望のアスペクト比や、厳密なセグメンテーションマスクからバウンディングボックスまで幅広い信号が含まれます。

1. はじめに

テキストから画像を生成するモデルは、”破壊的技術”として台頭し、テキストプロンプトから高品質で多様な画像を生成する能力を示しています。この分野では、現在、拡散モデルが最先端技術として確立されています (Saharia et al., 2022b; Ramesh et al., 2022; Rombach et al., 2022; Croitoru et al., 2022)。この進展は、デジタルコンテンツの作成方法を変革する大きな可能性を秘めていますが、テキストから画像を生成するモデルを現実世界のアプリケーションに展開する際には課題が残っています。その主要な課題は、生成されるコンテンツを直感的にユーザーが制御できるようにすることの難しさです。

現在、拡散モデルの制御性は以下の 2 つの方法で実現されています：

(i) モデルをゼロから学習させるか、既存の拡散モデルをタスクに合わせて微調整する方法（例：インペインティング、レイアウトから画像への学習など (Wang et al., 2022a; Ramesh et al., 2022; Rombach et al., 2022; Nichol et al., 2021; Avrahami et al., 2022b; Brooks et al., 2022; Wang et al., 2022b)）。しかし、モデルや学習データの規模が増大する中、この方法は微調整設定であっても膨大な計算リソースと長い開発期間を必要とする場合が多いです。

(ii) 事前学習済みのモデルを再利用し、制御可能な生成能力を追加する方法。これまで、これらの方法は特定のタスクに集中し、対象に特化した手法が設計されてきました（例：画像内のオブジェクトの置き換え、スタイルの操作、レイアウトの制御など (Tumanyan et al., 2022; Hertz et al., 2022; Avrahami et al., 2022a)）。

本研究の目標は、事前学習された（参照）拡散モデルを制御可能な画像生成に適応させる柔軟性を大幅に向上させる新しい統一フレームワークである **MultiDiffusion** を設計することです。MultiDiffusion の基本的なアイデアは、いくつかの参照拡散生成プロセスを共有パラメータや制約によって結びつけた新しい生成プロセスを定義することにあります。具体的には、参照拡散モデルを生成画像内の異なる領域に適用し、それぞれの領域についてノイズ除去のサンプリングステップを予測します。一方で、MultiDiffusion は、これらの異なるステップを最小二乗法による最適解を用いて統一的に調整することで、グローバルなノイズ除去サンプリングステップを実行します。

例えば、正方形の画像で学習された参照拡散モデルを用いて、任意のアスペクト比の画像を生成するタスクを考えます（図 2 参照）。各ノイズ除去ステップで、MultiDiffusion は参照モデルから提供される正方形のクロップ領域すべてのノイズ除去方向を統合し、近接するクロップ領域が共通のピクセルを共有しているという制約の下、それらすべての方向に可能な限り忠実であろうとします。直感的には、各クロップ領域が参照モデルからの真のサンプルとなるよう促します。各クロップ領域が異なるノイズ除去方向に引っ張られる可能性がある一方で、本フレームワークでは統一されたノイズ除去ステップを生成するため、高品質でシームレスな画像が作られます。

MultiDiffusion を使用することで、事前学習されたテキストから画像への参照モデルを、希望する解像度やアスペクト比の画像を生成するタスクや、大まかな領域ベースのテキストプロンプトを使用した画像生成といったさまざまなアプリケーションに活用することができます（図 1 参照）。特に、本フレームワークはこれらのタスクを共通の生成プロセスを用いて同時に解決することを可能にします。

関連するベースライン手法と比較した結果、本手法はこれらのタスク専用に学習された手法と比較しても、最先端の制御された生成品質を実現できることがわかりました。さらに、本手法は計算コストを追加することなく効率的に動作します。

1 関連研究

1.1 拡散モデル

拡散モデル（Sohl-Dickstein ら, 2015; Croitoru ら, 2022; Dhariwal & Nichol, 2021; Ho ら, 2020; Nichol & Dhariwal, 2021）は、データ分布 q を近似し、簡単にサンプリングできる生成確率モデルの一種です。具体的には、これらのモデルはガウスノイズ入力 $I_T \sim \mathcal{N}(0, I)$ を取り、段階的なノイズ除去ステップを通じてサンプル I_0 に変換します。このサンプルは q に従った分布となるべきです。ノイズ除去ステップの数や変換のパラメータ化は研究ごとに異なります（Sohl-Dickstein ら, 2015; Ho ら, 2020; Song ら, 2020; Lu ら, 2022a;b; Liu ら, 2022）。近年、拡散モデルは複雑な分布を学習し、多様で高品質なサンプルを生成する能力により、最先端の生成モデルとして位置付けられています。これらのモデルは、画像（Dhariwal & Nichol, 2021; Nichol & Dhariwal, 2021; Saharia ら, 2022b; Ramesh ら, 2022; Rombach ら, 2022）、動画（Ho ら, 2022; Singer ら, 2022）、3D シーン（Muller ら, 2022）、動作シーケンス（Yuan ら, 2022; Tevet ら, 2022）など、さまざまな領域で成功を収めています。

1.2 拡散モデルを用いた制御可能な生成

拡散モデルは、誘導入力チャネル（例: セマンティックレイアウト、カテゴリラベル）を用いて訓練することで、条件付き画像生成を成功させています（Ramesh ら, 2021; Saharia ら, 2022c;a; Wang ら, 2022a; Preechakul ら, 2022; Ho & Salimans, 2022）。最近のテキストから画像への拡散モデルは、画期的な生成能力を示す条件付き拡散モデルの最も顕著な例です（Nichol ら, 2021; Saharia ら, 2022b; Ramesh ら, 2022; Nichol ら, 2021; Rombach ら, 2022; Sheynin ら, 2022）。しかし、これらのモデルが生成コンテンツに対して提供する制御はわずかであり、主に入力テキストを介して達成されます。

近年、より広範かつ優れたユーザー制御性を実現する方法が多数提案されています。既存の方法は大きく 2 つに分類できます。(i) モデルに追加の誘導信号を使用して明示的な制御を組み込む方法（Avrahami ら, 2022b; Rombach ら, 2022; Brooks ら, 2022）。しかし、これらの手法は厳選されたデータセットでのコストが高い大規模な訓練を必要とします。(ii) 既存のモデルの生成プロセスを操作することで、暗黙的に生成コンテンツを制御する方法（Kwon & Ye, 2022; Meng ら, 2021; Tumanyan ら, 2022; Hertz ら, 2022; Avrahami ら, 2022c; Choi ら, 2021; Mokady ら, 2022; Couairon ら, 2022; Kong ら, 2023; Kwon ら, 2022）や、軽量なモデルの微調整を行う方法（Ruiz ら, 2022; Kawar ら, 2022; Kim ら, 2022; Valevski ら, 2022）があります。

Avrahami らは、微調整を必要としない画像インペインティング手法を設計しました（Avrahami ら, 2022a;c）。最近の研究（Tumanyan ら, 2022; Hertz ら, 2022）は、事前学習されたモデルの内部特徴やアーキテクチャ特性に関する洞察に基づいて、画像編集技術を設計しています。本研究では、事前学習された拡散モデルの生成プロセスを操作し、訓練や微調整を必要としません。しかし、明確な目標がない特定のアプリケーションを対象とした既存の研究とは対照的に、より一般的なアプローチを提案し、異なるユーザー制御入力をより体系的に統一することを可能にします。

2 方法論

本研究では、参照モデルとして機能する事前学習された拡散モデル Φ を考えます：

$$\Phi : I \times Y \to I$$

ここで、$I = R^{H \times W \times C}$ は画像空間、Y は条件空間（例えば、$y \in Y$ はテキストプロンプト）です。$I_T \sim P_I$ で初期化し、P_I は独立同分布のガウス分布を表し、条件 $y \in Y$ を設定したとき、拡散モデルは以下のように画像のシーケンスを構築します：

$$I_T, I_{T-1}, \ldots, I_0 \quad s.t. \quad I_{t-1} = \Phi(I_t|y)$$

この過程で、ノイズ画像 I_T を徐々にクリーン画像 I_0 に変換します。

2.1 MultiDiffusion

本研究の目標は、学習や微調整を行うことなく、Φ を活用して、異なる画像空間 $J = R^{H_0 \times W_0 \times C}$ および条件空間 Z で画像を生成することです。そのために、MultiDiffusion プロセスを次のように定義します：

$$\Psi : J \times Z \to J$$

MultiDiffusion は、拡散プロセスと同様に、初期ノイズ入力 $J_T \sim P_J$（ここで P_J は J 上のノイズ分布）から始まり、以下の画像列を生成します：

$$J_T, J_{T-1}, \ldots, J_0 \quad s.t. \quad J_{t-1} = \Psi(J_t|z)$$

本研究の基本的なアイデアは、Φ との整合性をできる限り維持する形で Ψ を定義することです。具体的には、ターゲット画像空間と参照画像空間の間のマッピング $F_i : J \to I$ と、条件空間間の対応するマッピング $\lambda_i : Z \to Y$ を定義します（ここで $i \in [n] = \{1, \ldots, n\}$）。これらのマッピングはアプリケーション依存であり、詳細は後述するセクション 4 で説明します。

MultiDiffuser ステップ $J_{t-1} = \Psi(J_t|z)$ を、参照モデル Φ が条件下で異なる画像領域に適用されたときのノイズ除去ステップ $\Phi(I_t^i|y_i)$ に可能な限り従わせることを目指します。すなわち：

$$I_t^i = F_i(J_t), \quad y_i = \lambda_i(z)$$

形式的には、新しいプロセスは次の最適化問題を解くことで定義されます：

$$\Psi(J_t|z) = \arg\min_{J \in J} \mathcal{L}_{FTD}(J|J_t, z)$$

ここで、

$$\mathcal{L}_{FTD}(J|J_t, z) = \sum_{i=1}^{n} W_i \otimes \left[F_i(J) - \Phi(I_t^i|y_i)\right]^2$$

$W_i \in R_+^{H \times W}$ はピクセルごとの重み、$\otimes$ はハダマード積を表します。

直感的に、FTD 損失は、参照モデル Φ が異なる領域 $F_i(J_t)$ に基づいて提案するノイズ除去サンプリングステップ $\Phi(I_t^i|y_i)$ を最小二乗の意味で調整します。図 2 は MultiDiffuser の 1 ステップを示しており、アルゴリズム 2 は MultiDiffusion サンプリングプロセスを要約しています。

3 閉形式の公式と MultiDiffusion の特性

3.1 閉形式の公式

本論文で示されたアプリケーションでは、F_i は直接的なピクセルサンプリング（例えば、画像 J_t の一部を切り抜く操作）から構成されています。この場合、式（4）は二次最小二乗（Least-Squares, LS）の形式を取り、最小化された J の各ピクセルは、すべての拡散サンプル更新の加重平均となります。具体的には次のように表されます：

$$\Psi(J_t|z) = \sum_{i=1}^{n} F_i^{-1}(W_i) \left(\sum_{j=1}^{n} F_j^{-1}(W_j) \right)^{-1} \otimes F_i^{-1}(\Phi(I_t^i|y_i))$$

ここで、W_i は各ピクセルの重み、$\otimes$ はハダマード積を表します。

3.2 MultiDiffusion の特性

式（3）で定義された Ψ の主な動機は、以下の観察に基づいています。確率分布 P_J を次の条件を満たすよう選択すると：

$$F_i(J_T) \sim P_I, \quad \forall i \in [n]$$

そして、式（3）の定義に基づき $J_{t-1} = \Psi(J_t|z)$ を計算し、FTC 損失 $L_{FTD}(J_{t-1}|J_t, z) = 0$ に到達した場合、以下が成り立ちます：

$$I_{t-1}^i = F_i(J_t) = \Phi(I_t^i|y_i)$$

すなわち、すべての $i \in [n]$ について、I_t^i は拡散シーケンスであり、I_0^i は Φ によって定義される画像空間 I 上の分布に従います。

3.3 命題 3.1

確率分布 P_J が式（6）を満たし、全ステップ $T, T-1, \ldots, 0$ において FTD コスト（式（4））がゼロに最小化される場合、画像 $I_t^i = F_i(J_t)$ は Φ の拡散経路を再現します。特に、$F_i(J_0)$（$i \in [n]$）は参照拡散モデル Φ からのサンプルと同一分布に従います。

3.4 命題の意義

この命題が持つ意味は非常に広範です。単一の参照拡散プロセスを使用することで、モデルを再訓練することなく、異なる画像生成シナリオに柔軟に適応することが可能になります。それでいて、参照拡散モデルとの一貫性を保つことができます。

次に、このフレームワークを具体化し、「拡散経路に従う（Follow-the-Diffusion-Paths）」アプローチのいくつかの応用について概説します。

4 応用例

4.1 パノラマ生成

最初の応用例として、本フレームワークを使用して、H0 H、W0 H を満たす画像空間 J における拡散モデルを定義します。これは、画像空間 I において動作する訓練済みモデル Φ に基づいています。この場合、$Z = Y$（つまり、与えられたテキストプロンプトに対してパノラマ画像を生成）とし、$F_i(J) \in I$ は画像 J の $H \times W$ の切り抜き部分、また $z = \lambda_i(z)$ とします。画像 J をカバーする n 個の切り抜き部分を考えます。$W_i = 1$ と設定すると、以下の式を得ます。

$$\Psi(J_t, z) = \arg\min_{J \in J} \sum_{i=1}^{n} \|F_i(J) - \Phi(F_i(J), z)\|^2 \tag{1}$$

これは最小二乗問題であり、その解は式（5）に基づいて解析的に計算されます。実装の詳細については付録 B.1 を参照してください。

セクション 3 で議論したように、MultiDiffusion は参照モデル Φ によって提供される複数の拡散経路を調整します。図 3 に示すように、$H \times 4W$ のパノラマを考えます。図 3(a) では、Φ を 4 つの非重複切り抜き部分に対して独立に適用した場合の生成結果を示します。予想通り、各切り抜き間で一貫性がありません。これはモデルからのランダムサンプルを 4 つ取得した結果に相当します。一方、同じ初期ノイズから開始し、本生成プロセス（式（7））を適用すると、これらの初期的に無関係な拡散経路を融合させ、高品質で一貫性のあるパノラマを生成できます（図 3(b)）。

4.2 領域ベースのテキストから画像生成

領域マスクの集合 $\{M_i\}_{i=1}^n \subset \{0,1\}^{H \times W}$ と、対応するテキストプロンプトの集合 $\{y_i\}_{i=1}^n \subset Y^n$ が与えられたとき、それぞれの領域で望ましい内容を表現する高品質な画像 $I \in I$ を生成することが目標です。すなわち、画像セグメント $I \otimes M_i$ は y_i を反映するべきです。

この目的において、式（2）に戻り、MultiDiffusion プロセスを条件空間 $Z = Y^n$（すなわち、$z = (y_1, \ldots, y_n)$）およびターゲット画像空間 $J = I$（参照画像空間と同一）に対して定義します。

$$\Psi : I \times Y^n \to I \tag{2}$$

さらに、領域選択マップは以下で定義されます。

$$F_i(I) = I \tag{3}$$

ピクセルの重みはマスクに基づいて設定されます。

$$W_i = M_i \tag{4}$$

Ψ ステップは以下の最小二乗問題の解として定義されます。

$$\Psi(J_t, z) = \arg\min_{J \in I} \sum_{i=1}^{n} \|M_i \otimes (J - \Phi(J_t | y_i))\|^2 \tag{5}$$

この最小二乗問題の解は解析的に計算されます。各ステップで、与えられたプロンプトに基づいて事前学習済みの拡散を適用し、複数の拡散方向 $\Phi(J_t|y_i)$ を得ます。式（5）に基づき、J_t 内の各ピクセルが含まれる領域 M_i に関連付けられた（平均化された）方向に従うように促します。

4.3 高い忠実度での厳密なマスクへの適合

ユーザーによって提供される厳密なマスクに対して高い忠実度を達成する手法をさらにサポートします（図5参照）。レイアウトは拡散プロセスの初期段階で決定されることに注目し、拡散モデル $\Phi(J_t|y_i)$ が早い段階で領域 M_i に焦点を当てて望ましいレイアウトに一致するよう促し、次に画像全体の文脈を考慮して調和のとれた結果を達成することを目指します。

この目的のために、マップ F_i に時間依存性を組み込み、ブートストラッピングフェーズを導入します。これは以下の式で定義されます。

$$F_i(J_t, t) = \{ J_t, if t \leq T_{init} M_i \otimes J_t + (1 - M_i) \otimes S_t, otherwise \tag{6}$$

ここで、T_{init} はブートストラッピングの停止ステップパラメータを表し、S_t は一定の色を持つランダムな画像で、背景として機能します（実装の詳細については付録 B.2 を参照してください）。

セクション 5.2 で、このブートストラッピング手法の効率性を実証します。我々は T_{init} を生成プロセスの 20%（すなわち、$T_{init} = 800$）に設定しました。

5 結果

セクション 4 で説明した各タスクに適用した際の本手法を徹底的に評価しました。すべての実験において、Stable Diffusion（Rombach et al., 2022）を使用しました。この拡散プロセスは潜在空間 $I = R^{64\times64\times4}$ 上で定義され、高解像度の自然画像 $[0,1]^{512\times512\times3}$ を再構築するデコーダが訓練されています。同様に、MultiDiffusion プロセス Ψ は潜在空間 $J = R^{H'\times W'\times4}$ 上で定義され、デコーダを使用してターゲット画像空間 $[0,1]^{8H_0\times8W_0\times3}$ で結果を生成します。

Stable Diffusion の公開実装（von Platen et al., 2022）の v2 事前学習モデルを使用しました。この実装は HuggingFace によって提供されています。

5.1 5.1 パノラマ生成

本手法をテキストからパノラマを生成するタスク（セクション 4.1）で評価するため、元の訓練解像度の 9 倍の幅となる 512 × 4608 の多様なパノラマを生成しました。テキストから任意のアスペクト比の画像を直接生成する方法が存在しないため、以下の 2 つのベースライン手法と比較しました: (i) Blended Latent Diffusion (BLD) (Avrahami et al., 2022a)（Stable Diffusion (Rombach et al., 2022) と組み合わせ）、および Stable Inpainting (SI) (Rombach et al., 2022)、後者は大規模データで微調整されたインペインティング用モデルです。両ベースライン手法では、中心の画像（入力テキストでモデル Φ からサンプリングされたもの）から徐々に右方向および左方向へ拡張してパノラマ画像を生成します。

図 4 は、本手法とこれらベースライン手法を用いた生成結果を示しています。図から分かるように、両ベースライン手法では重複領域に目立つ継ぎ目や不連続性が生じ、中心画像から離れるほど視覚的な品質が劣化することがよく見られます。これは反復的な生成プロセスの影響で予測される結果です。BLD はしばしば内容が繰り返される（例: スキーヤーの例）、一方で SI は画像の左部分と右部分で明らかな視覚的な違いを生じさせます。対照的に、本フレームワークはすべてのクロップの拡散経路を組み合わせることでパノラマ画像を同時に「サンプリング」し、シームレスで高品質な画像を生成します。追加の比較結果は付録 10 に示されています。

これらの観察を定量化するため、Frechet Inception Distance (FID) (Parmar et al., 2022) を使用して、パノラマ画像から得られる 512 × 512 のクロップ画像の分布と、参照モデル Φ が生成した画像の分布との距離を測定しました。具体的には、特定のテキストプロンプトに対して、Φ から異なる 512 × 512 画像を N サンプル取得し、それらを参照データセットとしました。ベースライン手法および本手法では、N のパノラマ画像を生成し、それぞれのサンプルからランダムに 512 × 512 のクロップを取得して生成データセットとし、FID を計算しました。

結果の品質をさらに評価するため、以下の 2 つの CLIP ベーススコアを測定しました: (i) テキストと画像の CLIP スコア（Radford et al., 2021）。これはテキストプロンプトと画像埋め込み間のコサイン類似度で評価されます。(ii) CLIP Aesthetic スコア（Schuhmann et al., 2022）。これは CLIP 上で学習した線形推定器によって画像の美的品質を予測します。

$N = 2000$ サンプルを使用し、8 つの異なるテキスト条件でこの評価を繰り返しました。表 1 には、本手法およびベースライン手法に対する FID および CLIP スコアの平均値と標準偏差を示します。また、Φ から独立して取得したサンプル画像セットのスコアも報告し、これをベースラインとして使用しました。結果として、本手法がすべての指標で既存のベースラインを上回ることが確認されました。

5.2　5.2 領域ベースのテキストから画像生成

本手法の領域ベースの定式化（セクション 4.2）は、厳密なマスク作成の負担を軽減することで、初心者ユーザーに柔軟なコンテンツ作成の手段を提供します。図 1、図 7、および図 8 に示すように、本手法はバウンディングボックス領域ガイダンスのみを与えられた場合でも、テキスト記述に従った多様で高品質なサンプルを生成します。図 7 では、異なる入力ノイズから生成を開始することで、同じ空間的制約に従いながらも、異なるスケールや外観でオブジェクトを描写する多様なサンプルを生成できることが分かります。特筆すべきは、本手法がすべての領域の制御を統一された生成プロセスに統合しているため、背景のぼかし、影、反射といった複雑なシーン効果を一貫して生成できる点です。追加の結果は付録に含まれています。

我々は、本手法を Make-A-Scene (Gafni et al., 2022) および同時研究である SpaText (Avrahami et al., 2022b) と比較しました。これらのベースライン手法は、このタスクのために大規模な学習を行っています。これらのモデルは公開されていないため、提供されている例に基づいて質的に比較を行いました。

さらに、BLD (Avrahami et al., 2022a) の適応版をベースラインとして考慮しました。セクション 5.1 と同様に、背景を最初に生成し、その後、前景オブジェクトを順次生成する自動回帰的な方法を適用しています。

図 5 に示すように、本手法は空間的制約を満たし、一貫性のある画像を生成します。また、(Avrahami et al., 2022b) に匹敵する質的な結果を示します。一方、BLD (Avrahami et al., 2022a) に基づく自動回帰的アプローチは、一貫性のない画像や不自然なシーン（例：「バスルーム」の例ではシンクの配置が不適切）を生成することが多いです。ベースラインとの追加比較は付録に含まれています。

本手法の性能を定量的に評価するために、COCO データセット (Lin et al., 2014) を使用しました。このデータセットには、画像全体のテキストキャプションと各オブジェクトのインスタンスマスクが含まれています。我々は、前景オブジェクトが 2 から 4 個で構成され、人物を含まず、マスクが画像の 5%未満を占める例をフィルタリングすることで、検証セットのサブセットを取得しました。これにより、1,000 の多様なサンプルが得られました。(Avrahami et al., 2022b) に従い、各前景領域に対して「a label」というテキストプロンプトを、背景を記述するためのプロンプトとして画像全体のキャプションを使用しました。

生成された画像について、既製のセグメンテーションモデル (Cheng et al., 2022) を使用し、生成結果とグラウンドトゥルースセグメンテーションとの交差オーバーユニオン（IoU）を測定しました。表 2 は、我々の手法および前述のベースライン手法に対する性能を示しています。上限値として、セット内の元画像との IoU も報告しました。我々の手法は、既存のベースライン（SI (Rombach et al., 2022) および BLD (Avrahami et al., 2022a)）を上回る結果を示しています。追加の質的な例は付録に含まれています。

最後に、我々のブートストラップ段階（式 9）のアブレーションを図 6 で質的に、表 2 で定量的に示しました。ブートストラップがない場合でも、本手法はマスク領域内に目的のオブジェクトを生成しますが、ブートストラップ段階により、マスクにより適合した結果を得られることが確認されました。

6　6. 考察と結論

制御可能な生成は、テキストから画像を生成する拡散モデルにおける主要な未解決課題の 1 つです。本研究では、事前学習された固定された拡散モデルの上に新たな生成プロセスを定義するという、根本的に新しい方向性からこの課題に取り組むことを提案しました。このアプローチは、従来の研究と比較して以下のような重要な利点を持ちます。(i) さらなるトレーニングや微調整を必要としない、(ii) さまざまな生成タスクに適用可能である、(iii) 多くのタスクにおいて閉形式で解ける最適化問題を生成プロセスとして提供するため、効率的に計算でき、目標関数のグローバル最適解に収束することを保証します。

制限事項としては、本手法が参照拡散モデルの生成事前分布に大きく依存している点が挙げられます。すなわち、結果の品質はモデルによって提供される拡散パスに依存します。したがって、参照モデルが「悪い」パス（例えば、不適切なシードや偏ったテキストプロンプト）を選択した場合、結果に影響を及ぼす可能性があります。この問題に対しては、フレームワークにさらなる制約を導入する（セクション 4.2 のブートストラッピング）、あるいはプロンプトエンジニアリング（図 9）を用いることで、ある程度緩和できる場合があります。

我々はフレームワークを徹底的に評価し、特定のタスクのために特別に訓練された手法と比較しても、最先端の結果を示しました。本研究が、事前学習された拡散モデルの力をより原理的な方法で活用するための今後の研究を促進することを期待しています。

例えば、MultiDiffusion をより一般的な最適化問題に拡張する方法が考えられます。

$$\Psi(J_t|z) = \arg\min_{J\in\mathcal{J}} \left[L_{FTD}(J|J_t,z) + L_0(J,J_t,z)\right] \quad s.t. \quad J \in \mathcal{C}(J_t,z) \tag{7}$$

ここで、L_0 はコスト関数、$\mathcal{C}$ は MultiDiffusion プロセスを制御するために、他の事前分布や設計制約を組み込む一連の（ハード）制約を表します。このアプローチは、MultiDiffusion プロセスを設計する際にさらなる自由度を提供します。

参考文献

Avrahami, O., Fried, O., and Lischinski, D. Blended latent diffusion. arXiv preprint arXiv:2206.02779, 2022a.

Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., and Yin, X. Spatext: Spatio-textual representation for controllable image generation. arXiv preprint arXiv:2211.14305, 2022b.

Avrahami, O., Lischinski, D., and Fried, O. Blended diffusion for text-driven editing of natural images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18208–18218, 2022c.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. November 2022.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938, 2021.

Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427, 2022.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. arXiv preprint arXiv:2209.04747, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 2021.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. In European Conference on Computer Vision (ECCV), 2022.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. arXiv preprint arXiv:2210.09276, 2022.

Kim, G., Kwon, T., and Ye, J. C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2426–2435, 2022.

Kong, C., Jeon, D., Kwon, O., and Kwak, N. Leveraging off-the-shelf diffusion model for multi-attribute fashion image manipulation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023.

Kwon, G. and Ye, J. C. Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264, 2022.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In European conference on computer vision, pp. 740–755. Springer, 2014.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778, 2022.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. arXiv preprint arXiv:2206.00927, 2022a. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095, 2022b.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073, 2021.

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. Null-text inversion for editing real

images using guided diffusion models. arXiv preprint arXiv:2211.09794, 2022.
Muller, N., Siddiqui, Y., Porzi, L., Rota Bul ´ o, S., Kontschieder, P., and Nießner, M. Diffrf: Rendering-guided 3d radiance field diffusion. arxiv, 2022.
Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021.
Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In International Conference on Machine Learning, pp. 8162–8171. PMLR, 2021.
Parmar, G., Zhang, R., and Zhu, J.-Y. On aliased resizing and surprising subtleties in gan evaluation. In CVPR, 2022.
Preechakul, K., Chatthee, N., Wizadwongsa, S., and Suwajanakorn, S. Diffusion autoencoders: Toward a meaningful and decodable representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10619–10629, 2022.
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. In International Conference on Machine Learning, pp. 8748–8763. PMLR, 2021.
Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In International Conference on Machine Learning, pp. 8821–8831. PMLR, 2021.
Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.
Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242, 2022.
Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In ACM SIGGRAPH 2022 Conference Proceedings, 2022a.
Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487, 2022b.
Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022c.
Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models. ArXiv, abs/2210.08402, 2022.
Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., and Taigman, Y. Knn-diffusion: Image generation via large-scale retrieval. arXiv preprint arXiv:2204.02849, 2022.
Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., and Taigman, Y. Make-a-video: Text-to-video generation without text-video data, 2022.
Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pp. 2256–2265. PMLR, 2015.
Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.
Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A. H., and Cohen-Or, D. Human motion diffusion model. arXiv preprint arXiv:2209.14916, 2022.
Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation. arXiv preprint arXiv:2211.12572, 2022.
Valevski, D., Kalman, M., Matias, Y., and Leviathan, Y. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. arXiv preprint arXiv:2210.09477, 2022.
von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.
Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., and Wen, F. Pretraining is all you need for image-to-image translation. In arXiv, 2022a.
Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., and Li, H. Semantic image synthesis via

diffusion models. arXiv preprint arXiv:2207.00050, 2022b.
Yuan, Y., Song, J., Iqbal, U., Vahdat, A., and Kautz, J. Physdiff: Physics-guided human motion diffusion model. arXiv preprint arXiv:2212.02500, 2022.

A. Additional Results

以下のセクションでは、本論文で示したアプリケーションに関する追加結果および比較を提供します。

A.1. パノラマ生成

テキストからパノラマ生成タスク (Sec. 5.1) に関する追加結果および定性的な比較を提供します。図 10 では、本手法と Stable Inpainting (SI) (Rombach et al., 2022) および Blended Latent Diffusion (BLD) (Avrahami et al., 2022a) の比較を示しています。また、縦方向のパノラマ生成結果を図 12 左に示しています。

A.2. 領域ベースのテキストから画像生成

領域ベースの生成タスク (Sec. 4.2) に関する追加の定性的な結果および比較を図 12 および図 14 に示します。

A.3. COCO における領域ベースのテキストから画像生成

COCO の検証セットのサブセットに基づくサンプル結果および比較を図 13 に示します。この実験の詳細については Sec. 5.2 をご参照ください。

B. Additional Implementation Details

B.1. パノラマ (Sec. 4.1)

パノラマ生成の場合、マッピング F_i は全パノラマからの固定サイズのクロップとして定義されます。具体的には、空間解像度が $H_0 \times W_0$ のパノラマについて、Stable Diffusion の潜在空間でサイズ $H \times W$ （RGB 空間では 512×512 に変換）を持つオーバーラップするクロップを考えます。ここで、$F_i, ..., F_n$ は潜在空間においてステップサイズ $step = 8$ （RGB 空間では 64 ピクセル）のスライディングウィンドウとしてクロップを提供します。特に、クロップの総数 n は以下のように計算されます：

$$n = \frac{H_0 - 64}{step} \cdot \frac{W_0 - 64}{step}.$$

各クロップの拡散更新は並列（バッチ処理）で計算できます。その結果、参照拡散モデル Φ への呼び出し回数は合計 $T \cdot \frac{n}{b}$ 回となります。ここで、b はバッチサイズを示します。

B.2. ブートストラッピング (Sec. 4.2)

高い忠実度でタイトなマスクを保持したい場合（図 4 参照）、マップ F_i にブートストラッピングフェーズを導入します（Eq. 9 参照）。具体的には、各 S_t を以下の手順で事前計算します：

1. ランダムな定数 RGB 値を持つ画像 $I \in [0, 1]^{512 \times 512 \times 3}$ をランダム化します。
2. それを Stable Diffusion の事前学習エンコーダ E を用いて潜在空間 $S = E(I)$ にエンコードします。
3. 最後に、S_t を時刻ステップ t のノイズレベルにノイズ化します。すなわち、$S_t \sim N(\mu_t \cdot S, \sigma_t^2)$ となります。

ここで、μ_t および σ_t は拡散ノイズスケジューラ (Ho et al., 2020) を示します。