Universal Guidance for Diffusion Models

Arpit Bansal * 1, Hong-Min Chu * 1, Avi Schwarzschild 1, Soumyadip Sengupta 2, Micah Goldblum 3, Jonas Geiping 1, Tom Goldstein 1

概要

従来の拡散モデルは特定の形式の条件付け（主にテキスト）に適合するように訓練されており、他のモダリティで条件付けを行う場合には再訓練が必要です。本研究では、あらゆるガイダンスモダリティを用いて拡散モデルを制御できる**ユニバーサルガイダンスアルゴリズム**を提案します。このアルゴリズムにより、使用特有のコンポーネントを再訓練することなく、高品質な画像を生成することが可能になります。提案手法は、セグメンテーション、顔認識、物体検出、分類器信号などのガイダンス関数を用いた場合でも成功を収めました。

コード: GitHub リポジトリ
`github.com/arpitbansal297/Universal-GuidedDiffusion`

1 Introduction

拡散モデルはデジタルアートやグラフィック制作のための強力なツールです。その成功の多くは、出力を慎重に制御し、ユーザーごとの個別のニーズに合わせて結果をカスタマイズする能力に由来しています。今日のほとんどのモデルは条件付けによって制御されています。条件付けでは、拡散モデルがユーザーからの特定の入力モダリティ（記述的なテキスト、セグメンテーションマップ、クラスラベルなど）を受け入れるように設計されています。

条件付けは強力なツールですが、それによってモデルが単一の条件付けモダリティに縛られてしまいます。別のモダリティが必要になると、新しいモデルをゼロから再訓練する必要があります。しかし、トレーニングのコストが非常に高いため、大半のユーザーにとってこれは現実的ではありません。

モデル出力を制御するためのより柔軟なアプローチは、**ガイダンス**を使用することです。このアプローチでは、拡散モデルは汎用的な画像生成器として機能し、ユーザーの指示を直接理解する必要はありません。ユーザーは、このモデルを基準関数と組み合わせて使用します。この関数は、特定の基準が満たされているかどうかを測定します。例えば、生成された画像とユーザーが選択したテキスト記述との間の CLIP スコアを最小化するようにモデルを誘導することができます。

画像生成の各反復中に、反復体はガイダンス関数の勾配に沿ってわずかに移動し、最終的な生成画像がユーザーの基準を満たすようになります。

本論文では、任意の既存モデルや損失関数をガイダンスとして拡散モデルに利用できるガイダンス手法を研究します。ガイダンス関数は再訓練や改変なしで使用できるため、この形式のガイダンスは**ユニバーサル**であり、ほぼすべての目的に拡散モデルを適応させることが可能です。

ユーザーの観点から見ると、ガイダンスは条件付けよりも優れています。単一の拡散ネットワークが基盤モデルとして扱われ、多くのユースケース（一般的なものから特注のものまで）に対して普遍的なカバレッジを提供します。しかしながら、このアプローチは実現不可能だと広く信じられています。

初期の拡散モデルは分類器ガイダンス（Dhariwal Nichol, 2021）に依存していましたが、研究コミュニティはすぐに分類器を使用しない方式（Ho Salimans, 2022）へ移行しました。この方式では、モデルをクラスラベルに基づいて特定の固定されたオントロジーでゼロから訓練する必要があります（Nichol et al., 2021; Rombach et al., 2022; Bansal et al., 2022）。

ガイダンスの使用が難しい理由は、拡散サンプリングプロセスで使用されるノイズ画像と、ガイダンスモデルが訓練されるクリーンな画像との間のドメインシフトに起因します。このギャップが解

消されれば、ガイダンスは成功裏に実行できます。例えば、Nichol et al. (2021) は CLIP モデルをガイダンスとして成功裏に使用しましたが、ノイズの入った入力を使用して CLIP をゼロから再訓練する必要がありました。ノイズ再訓練はドメインギャップを埋めますが、非常に高額なコストとエンジニアリング負荷を伴います。

追加のコストを避けるために、我々はモデルではなくサンプリングスキームを変更することでこのギャップを埋める方法を研究します。

貢献の要約

本研究の貢献を以下のようにまとめます：

- **ユニバーサルガイダンスを可能にするアルゴリズムの提案:** 提案するサンプラーは、ノイズの多い潜在状態ではなく、除去されたノイズ画像に対してガイダンスモデルを評価します。これにより、従来のガイダンス手法を妨げていたドメインギャップを解消します。この戦略により、エンドユーザーは広範なガイダンスモダリティや複数のモダリティを同時に扱う柔軟性を得ます。基盤となる拡散モデルは固定されたままで、いかなるファインチューニングも不要です。
- **多様な制約への適用性の実証:** 提案手法の有効性を、分類器ラベル、人間のアイデンティティ、セグメンテーションマップ、物体検出器からの注釈、逆線形問題から生じる制約など、さまざまな制約に対して実証します。

2 Background

本節では、拡散モデルの基本フレームワークに関する最近の文献を簡単に概説します。その後、制御された画像生成の問題設定を定義し、関連する先行研究について議論します。

2.1 Diffusion Models

拡散モデルは強力な生成モデルであり、画像生成の分野で初めて導入された際にもその高い性能が証明されました（Song & Ermon, 2019; Ho et al., 2020）。このアプローチは、音声やテキスト生成など、多くの分野に成功裏に拡張されています（Kong et al., 2020; Huang et al., 2022; Austin et al., 2021; Li et al., 2022）。

本節では、異なる種類のモデルのニュアンスを説明する上で有用な、（無条件の）拡散を形式的に紹介します。拡散モデルは、T ステップの**前進プロセス**と T ステップの**逆プロセス**の組み合わせとして定義されます。概念的には、前進プロセスは、クリーンなデータ点 z_0 に対して異なる大きさのガウスノイズを段階的に加え、逆プロセスは、ノイズの入った入力を段階的にデノイズし、クリーンなデータ点を復元しようとします。

より具体的には、ノイズスケールを表すスカラーの配列 $\{\alpha_t\}_{t=1}^T$ と初期のクリーンデータ点 z_0 が与えられた場合、前進プロセスの t ステップを z_0 に適用すると、ノイズデータ点 z_t が得られます。

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

拡散モデルは学習されたデノイズネットワーク ϵ_θ であり、任意のペア (z_0, t) および任意のサンプル ϵ に対して以下を満たすように訓練されます：

$$\epsilon_\theta(z_t, t) \approx \epsilon = \frac{z_t - \sqrt{\alpha_t} z_0}{\sqrt{1-\alpha_t}}.$$

逆プロセスは $q(z_{t-1}|z_t, z_0)$ という形を取り、さまざまな詳細な定義が存在します。ここで $q(\cdot|\cdot)$ は一般的にガウス分布としてパラメータ化されます。異なる研究では、未知の $q(z_{t-1}|z_t, z_0)$ を用いたサンプリングの近似についても研究されています。たとえば、Denoising Diffusion Implicit Model（DDIM）(Song et al., 2021a) では、予測されるクリーンデータ点 $\hat{z}_0$ を以下のように計算します：

$$\hat{z}_0 = \frac{z_t - (\sqrt{1-\alpha_t})\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}},$$

その後、未知の z_0 を $\hat{z}_0$ に置き換えることで、$q(z_{t-1}|z_t, \hat{z}_0)$ から z_{t-1} をサンプリングします。

一方で、個々のサンプリング手法の詳細は異なるものの、すべてのサンプリング手法は現在のサンプル z_t、現在のタイムステップ t、および予測されたノイズ $\hat{\epsilon}$ に基づいて z_{t-1} を生成します。

表記の簡略化のため、サンプリング手法を抽象化する関数 $S(\cdot, \cdot, \cdot)$ を定義します。この場合、サンプリングは以下のように表されます：

$$z_{t-1} = S(z_t, \hat{\epsilon}, t).$$

2.2 制御された画像生成

本論文では、さまざまな制約を用いた制御された画像生成に焦点を当てます。微分可能なガイダンス関数 f（例えば、CLIP 特徴抽出器やセグメンテーションネットワーク）を考えます。画像 x に適用すると、ベクトル $c = f(x)$ が得られます。また、2 つのベクトル c と c' の近さを測定する関数 $\ell(\cdot, \cdot)$ を考えます。特定の c の選択をプロンプトと呼び、対応する制約は c, ℓ, f に基づいて $\ell(c, f(z)) \approx 0$ と形式化されます。我々の目的は、この制約を満たす画像分布からサンプル z を生成することです。簡単に言うと、プロンプトに一致する分布内の画像を生成したいということです。

制御された生成型拡散モデルを研究した先行研究は、主に 2 つのカテゴリに分類されます。第 1 のカテゴリを条件付き画像生成（Conditional Image Generation）、第 2 のカテゴリをガイド付き画像生成（Guided Image Generation）と呼びます。次に、それぞれのカテゴリの特性を説明し、本研究の位置づけを明確にします。

2.2.1 条件付き画像生成

このカテゴリの手法は、プロンプトを追加入力として受け取る新しい拡散モデルを訓練する必要があります（Ho & Salimans, 2022; Bansal et al., 2022; Nichol et al., 2021; Whang et al., 2022; Wang et al., 2022a）。例えば、Ho & Salimans (2022) は、クラスラベルをプロンプトとして使用し、非条件出力と条件付き出力の線形補間を通じて拡散モデルを訓練する分類器なしガイダンス（Classifier-Free Guidance）を提案しました。Bansal et al. (2022) は、ガイダンス関数が既知の線形劣化演算子である場合を研究し、線形逆問題を解く条件付きモデルを訓練しました。Nichol et al. (2021) は、CLIP（Radford et al., 2021）の表現を使用して、生成された画像とテキストプロンプト間の類似性を強制するテキスト条件付き画像生成に分類器なしガイダンスを拡張しました。

これらの方法は、さまざまなタイプの制約において成功を収めていますが、拡散モデルの再訓練が必要であるため、計算コストが非常に高くなります。

2.2.2 ガイド付き画像生成

このカテゴリの手法は、事前に学習された拡散モデルを基盤モデルとして使用しますが、ガイダンス関数からのフィードバックを利用してサンプリング手法を修正し、画像生成を誘導します。本研究の方法は、このカテゴリに該当します。過去の研究では、さまざまな制約や外部ガイダンス関数を用い

たガイド付き画像生成が検討されてきました (Dhariwal & Nichol, 2021; Kawar et al., 2022; Wang et al., 2022b; Chung et al., 2022a; Lugmayr et al., 2022; Chung et al., 2022b; Graikos et al., 2022)。

例えば、Dhariwal & Nichol (2021) は、異なるノイズスケールの画像に対して分類器を訓練し、これをガイダンス関数 f として使用する分類器ガイダンス (Classifier Guidance) を提案しました。そして、この分類器の勾配をサンプリングプロセスに含めました。しかし、ノイズが加えられた画像の分類器はドメイン特化型であり、一般にすぐに利用できるものではありません。この問題は本研究の手法では回避されています。Wang et al. (2022b) は外部ガイダンス関数が線形演算子であると仮定し、基盤モデルを使用して線形演算子の零空間内に存在する画像の成分を生成しました。しかし、この方法を非線形ガイダンス関数に拡張することは容易ではありません。Chung et al. (2022a) は一般的なガイダンス関数を検討し、期待されるノイズ除去画像上で計算されたガイダンス関数の勾配を使用してサンプリングプロセスを修正しました。しかし、著者らが提示した結果は非線形ブラーのような単純な非線形ガイダンス関数に限定されています。

本研究では、物体検出やセグメンテーションネットワークのような既存のガイダンス関数 f を使用し、拡散モデルを用いたガイド付き画像生成のための普遍的なガイダンスアルゴリズムを検討します。

3 Universal Guidance

我々は、拡散モデルの画像サンプリング手法を拡張し、市販の補助ネットワークからのガイダンスを組み込むためのガイダンスアルゴリズムを提案します。本アルゴリズムは、式 (3) によって得られる再構築されたクリーンな画像 $\hat{z}_0$ が、完全ではないものの、一般的なガイダンス関数に対して有益なフィードバックを提供し、画像生成を誘導するのに適しているという経験的観察に基づいています。セクション 3.1 では、この観察を活用し、汎用的なガイダンス関数を扱うために分類器ガイダンス (Dhariwal & Nichol, 2021) を拡張することで、前向きユニバーサルガイダンス (Forward Universal Guidance) を動機付けます。セクション 3.2 では、ガイダンス関数 f に基づく制約を満たすために生成画像を強制する補助的な後向きユニバーサルガイダンス (Backward Universal Guidance) を提案します。セクション 3.3 では、生成された画像の忠実度を経験的に改善するためのシンプルながら有用な自己再帰トリックについて議論します。

3.1 Forward Universal Guidance

外部ガイダンス関数 f と損失関数 ℓ からの情報を用いて生成をガイドするため、分類器ガイダンス (Dhariwal & Nichol, 2021) を任意の汎用的なガイダンス関数に対応させる方法を拡張することが直感的に考えられます。具体的には、クラスプロンプト c が与えられた場合、分類器ガイダンスは各サンプリングステップ $S(z_t, t)$ において、以下のように $_\theta(z_t, t)$ を置き換えることで、分類に基づくサンプリングを実行します：

$$_\theta(z_t, t) =_\theta (z_t, t) - \sqrt{1-\alpha_t}\nabla_{z_t} \log p(c|z_t).$$

定義として、$ce(\cdot,\cdot)$ をクロスエントロピー損失、f_{cl} を分類確率を出力するガイダンス関数とすると、式 (4) は以下のように書き換えられます：

$$\hat{\epsilon}_\theta(z_t, t) = \epsilon_\theta(z_t, t) + \sqrt{1-\alpha_t}\nabla_{z_t} l_{ce}(c, f_{cl}(z_t)). \tag{5}$$

しかし、f_{cl} と ce を任意の既製のガイダンス関数と損失関数に直接置き換えることは、実際には機能しません。これは、f がクリーンな画像で訓練されている可能性が高く、入力がノイズを含む場合には意味のあるガイダンスを提供できないためです。

この問題に対処するため、$\epsilon_\theta(z_t, t)$ がデータポイントに加えられたノイズを予測することを利用します。したがって、式 (3) により予測されるクリーンな画像 $\hat{z}_0$ を得ることができます。そこで、ガイダンスを予測されたクリーンなデータポイントに基づいて計算することを提案します：

$$\hat{\epsilon}_\theta(z_t, t) = \epsilon_\theta(z_t, t) + s(t) \cdot \nabla_{z_t} \ell(c, f(\hat{z}_0)), \tag{6}$$

ここで、$s(t)$ は各サンプリングステップのガイダンス強度を制御し、以下が成り立ちます：

$$\nabla_{z_t} \ell(c, f(\hat{z}_0)) = \nabla_{z_t} \ell \left(c, f \left(\frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}} \right) \right),$$

これは式 (3) に従います。この式 (6) を「前向きユニバーサルガイダンス（forward universal guidance)」、または単に「前向きガイダンス」と呼びます。

実際には、前向きガイダンスを適用することで、生成された画像がプロンプトに近づきつつ、生成の軌跡がデータの多様体上に維持されます。関連するアプローチとして、Chung ら (2022a) では、ガイダンスステップを $E[z_0|z_t]$ に基づいて計算する方法も研究されています。このアプローチは、スコアベース生成フレームワーク (Song et al., 2021b) に着想を得たものですが、異なる更新方法をもたらしました。

3.2 Backward Universal Guidance

前向きガイダンスの欠点として、画像の「現実感」を維持することを過度に優先し、与えられたプロンプトとの一致が不十分になる場合があることを観察しました（セクション 4.2 を参照)。単にガイダンス強度 $s(t)$ を増加させる方法は最適ではなく、これにより画像が多様体から外れる速度が増し、デノイザーがそれを補正するのが困難になることがしばしばあります。

この問題に対処するため、前向きガイダンスを補完する「後ろ向きユニバーサルガイダンス（backward universal guidance)」、または単に「後ろ向きガイダンス」を提案します。後ろ向きガイダンスの主なアイデアは、クリーンな画像 $\hat{z}_0$ を基にプロンプトに最も一致するよう最適化し、その修正を線形的にノイズ画像空間に戻すことです。具体的には、$\nabla_{z_t} \ell(c, f(\hat{z}_0))$ を直接計算する代わりに、クリーンデータ空間での修正 Δz_0 を次のように計算します：

$$\Delta z_0 = \arg\min_{\Delta} \ell(c, f(\hat{z}_0 + \Delta)). \tag{7}$$

実験的には、$\Delta = 0$ を初期値とし、m ステップの勾配降下法を用いて式 (7) を解きます。$\hat{z}_0 + \Delta z_0$ が $\ell(c, f(z))$ を直接最小化するため、Δz_0 は制約を最適に反映するクリーンデータ空間での変化量となります。その後、Δz_0 をノイズデータ空間 z_t に戻すために、以下を満たす誘導デノイジング予測 $\tilde{\epsilon}$ を計算します：

$$z_t = \sqrt{\alpha_t}(\hat{z}_0 + \Delta z_0) + \sqrt{1-\alpha_t}\tilde{\epsilon}. \tag{8}$$

式 (3) を再利用すると、$\tilde{\epsilon}$ は元のデノイジング予測 $\epsilon_\theta(z_t, t)$ に対する拡張として次のように書けます：

$$\tilde{\epsilon} = \epsilon_\theta(z_t, t) - \sqrt{\alpha_t/(1-\alpha_t)}\Delta z_0. \tag{9}$$

前向きガイダンスと比較すると、後ろ向きガイダンス（式 (9)）は生成画像が与えられたプロンプトに一致するための最適な方向を提供し、制約の強制を優先します。さらに、式 (7) の勾配ステップの計算は前向きガイダンス（式 (6)）よりも計算コストが低いため、複数の勾配ステップで式 (7) を解くことが可能であり、プロンプトとの一致をさらに向上させることができます。

なお、「前向き」および「後ろ向き」という名前は、オイラー法における前進オイラー法と後退オイラー法に類似しています。

3.3 3.3. ステップごとの自己再帰 (Per-step Self-recurrence)

ユニバーサルガイダンスを標準的な生成パイプラインに適用すると、しばしばアーティファクトや奇妙な振る舞いが見られ、自然画像とは明確に区別される場合があります。このような現象は、線形ガイダンス関数を研究した (Lugmayr et al., 2022; Wang et al., 2022b) においても観察されています。ガイダンスの強度 $s(t)$ を減少させて現実性を優先しようと試みましたが、ガイダンス制約を満たしつつ現実性を確保する「スイートスポット」は必ずしも存在するわけではありません。

この問題の原因として、ガイダンス関数が過剰な情報損失を引き起こした場合、生成画像が自然な画像サンプリング軌道から逸れる可能性があると推測されます。このため、ガイダンス方向が常に画像の現実性に関連するわけではありません。

(Lugmayr et al., 2022; Wang et al., 2022b) に着想を得て、この問題に対処するため、各ステップで自己再帰 (self-recurrence) を適用します。具体的には、$z_{t-1} = S(z_t, \hat{\epsilon}_t, t)$ がサンプリングされた後、ランダムなガウスノイズ $\epsilon' \sim \mathcal{N}(0, I)$ を z_{t-1} に再注入し、次のように z'_t を得ます：

$$z'_t = \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} \cdot z_{t-1} + \sqrt{1 - \frac{\alpha_t}{\alpha_{t-1}}} \cdot \epsilon'. \tag{10}$$

式 (10) により、z'_t は時刻 t に適したノイズスケールを持つことが保証されます。この自己再帰を k 回繰り返してから、ステップ $t-1$ のサンプリングを続けます。直感的には、自己再帰により、同じノイズスケールでデータ多様体の異なる領域を探索することが可能になり、ガイダンスと画像品質の両方を満たす解を見つけるための余地が広がります。

実験的には、この自己再帰により、適切なガイダンス強度 $s(t)$ を維持しつつ、与えられたプロンプトと一致する画像を生成することが可能であることを確認しました。自己再帰が生成画像の調和をどのように改善するかの例を図 2 に示します。

ユニバーサルガイダンスアルゴリズムの概要 前向きガイダンス、後ろ向きガイダンス、ステップごとの自己再帰から成るユニバーサルガイダンスアルゴリズムを Algorithm 1 にまとめます。このアルゴリズムは、単一のガイダンス関数を仮定していますが、複数の (f, ℓ) ペアを処理するように簡単に適応できます。また、前向きガイダンスと後ろ向きガイダンスの目的は同一である必要はなく、複数のガイダンス関数を同時に利用する異なる方法が可能です。

4 4. 実験 (Experiments)

本セクションでは、提案するユニバーサルガイダンスアルゴリズムを様々なガイダンス関数に対してテストした結果を示します。具体的には、Stable Diffusion (Rombach et al., 2022) と、ImageNet (Deng et al., 2009) 上で学習された純粋な無条件拡散モデルを使用します。Stable Diffusion はテキストプロンプトを追加入力として受け取ることでテキスト条件付き生成を行う拡散モデルですが、テキストプロンプトとして空文字列を指定することで無条件画像生成も可能です。

まず、Stable Diffusion に基づいた異なるガイダンス関数を用いた実験結果をセクション 4.1 で示し、次に ImageNet 拡散モデルに関する結果をセクション 4.2 で述べます

4.1 4.1. Stable Diffusion の結果 (Results for Stable Diffusion)

ここでは、Stable Diffusion を基盤モデルとして使用したガイダンス付き画像生成の結果を示します。使用したガイダンス関数には、CLIP フィーチャ抽出器 (Radford et al., 2021)、セグメンテーションネットワーク、顔認識ネットワーク、物体検出ネットワークが含まれます。Stable Diffusion を用いた

実験では、前方ガイダンスを適用するだけでプロンプトに一致する高品質な画像が生成されることが確認されたため、$m = 0$ に設定しました。

Stable Diffusion で前方ガイダンスを実行するには、式 (3) で計算された予測クリーン潜在変数を Stable Diffusion の画像デコーダに渡し、予測クリーン画像を取得します。それぞれのガイダンス関数に関する結果と実装の詳細は、対応するサブセクションで議論します。

CLIP ガイダンス (CLIP Guidance) CLIP (Radford et al., 2021) は、OpenAI によって開発された最先端のテキスト-画像類似度モデルです。本アルゴリズムをテキスト誘導画像生成に適用するため、CLIP の画像フィーチャ抽出器をガイダンス関数として使用します。テキストプロンプトによって生成される CLIP テキスト埋め込みと画像埋め込み間の負のコサイン類似度を計算する損失関数を構築しました。

$s(t) = 10\sqrt{1-\alpha_t}$、$k = 8$ に設定し、Stable Diffusion を無条件画像生成器として使用しました。様々なテキストプロンプトを用いて画像を生成しました。また、ユニバーサルガイダンスアルゴリズムの評価を深め、ガイダンスと条件付けを比較するため、Stable Diffusion を用いた従来のテキスト条件付き生成でも同一のプロンプトで画像を生成し、結果を図 3 にまとめました。

図 3 の結果は、本アルゴリズムが高品質な画像生成を誘導できることを示しており、生成された画像は専門的なテキスト条件付けモデルによって生成された画像と比較可能であることが確認されました。

セグメンテーションマップガイダンス (Segmentation Map Guidance) セグメンテーションマップをプロンプトとして使用した画像生成を誘導するため、MobileNetV3-Large (Howard et al., 2019) にセグメンテーションヘッドを追加したネットワークと、PyTorch (Paszke et al., 2019) で公開されている事前学習モデルを使用しました。セグメンテーションネットワークはピクセルごとの分類確率を出力するため、生成された画像の予測セグメンテーションと指定されたプロンプト間のピクセルごとのクロスエントロピー損失の合計を損失関数 $`\ell`\Gamma\Gamma\Gamma\Gamma\Gamma\Gamma\Gamma\Gamma\Gamma\Gamma\mathrm{s(t)} = 400 \cdot \sqrt{1-\alpha_t}$、$k = 10$ と設定しました。

実験では、異なる形状のオブジェクトを描いたセグメンテーションマップを新しいテキストプロンプトと組み合わせました。テキストプロンプトを Stable Diffusion の固定追加入力として使用してテキスト条件付きサンプリングを実行し、生成されたテキスト条件付き画像を与えられたセグメンテーションマップに一致するよう誘導しました。結果は図 4 に示されています。

図 4 より、生成された画像はオブジェクトと背景が明確に分離され、与えられたセグメンテーションマップとほぼ完全に一致していることが確認されます。さらに、生成されたオブジェクトと背景はそれぞれの記述テキスト（例：犬種や環境の説明）とも一致しており、画像全体の現実性も非常に高いことが分かります。

顔認識ガイダンス (Face Recognition Guidance) 特定の人物の顔に似せて画像生成を誘導するため、顔検出モジュールと顔認識モジュールを組み合わせたガイダンス関数を構築しました。このセットアップは入力顔画像から顔の属性埋め込みを生成します。顔検出モジュールには多タスクカスケード畳み込みネットワーク (MTCNN) (Zhang et al., 2016) を、顔認識モジュールには Facenet (Schroff et al., 2015) を使用しました。ガイダンス関数 f は検出された顔を切り抜いて顔属性埋め込みを出力し、損失関数 $`\ell`\Gamma\Gamma\Gamma\Gamma\Gamma\Gamma\Gamma\Gamma\mathrm{l}_1$ 損失を使用しました。

本アルゴリズムでガイダンス方向を計算する際には、Facenet のみをバックプロパゲーションし、MTCNN による非最大抑制 (Neubeck Van Gool, 2006) が非微分可能であるため、MTCNN が生成した顔切り抜きマスクをオラクル入力として扱いました。ここでは $s(t) = 20000 \cdot \sqrt{1-\alpha_t}$、$k = 2$ と設定しました。

顔ガイダンスとテキストプロンプトの組み合わせ (Face Guidance and Text Prompts) 顔ガイダンスとテキストプロンプトのさまざまな組み合わせを検討しました。セグメンテーションの場合と同様に、テキストプロンプトを Stable Diffusion の固定追加入力として使用し、顔のプロンプトが生成された画像の顔に似るよう、アルゴリズムでテキスト条件付き生成軌道を誘導しました。図 5 に示されている通り、指定された顔プロンプトの顔の特徴がほぼ完全に再現されています。さらに、背景、素材、スタイルのいずれかに関する説明テキストも正しく実現され、生成された顔と調和しています。

オブジェクト位置ガイダンス (Object Location Guidance) Stable Diffusion を用いて、オブジェクト検出ネットワークで画像生成を誘導する結果を提示します。この実験では、ResNet-50-FPN バックボーンを用いた Faster R-CNN (Ren et al., 2015; Li et al., 2021) をオブジェクト検出器として使用し、PyTorch で公開されている事前学習済みモデルを採用しました。バウンディングボックスとクラスラベルをオブジェクト位置プロンプトとして使用します。

損失関数 ‘ℓ‘ΓΓΓΓΓ3ΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓ1.ΓΓΓΓΓΓΓΓΓΓΓ($anchorclassification loss$)2.ΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓΓ

(1) と (2) はリージョンプロポーザルヘッドで計算され、(3) はリージョンクラス分類ヘッドで計算されます。標準的な R-CNN トレーニングと比較して、リージョンクラス分類ヘッドの追加バウンディングボックスアライメント損失を省略しました。この損失構成により、各位置プロンプトに対して正しいカテゴリのオブジェクトを生成できることが確認されました。$s(t) = 100 \cdot \sqrt{1-\alpha_t}$、$k = 3$ と設定しました。

テキストプロンプトとオブジェクト位置プロンプトのさまざまな組み合わせについても実験を行い、テキストプロンプトを Stable Diffusion の固定条件付けとして使用しました。我々の提案するガイダンスアルゴリズムを用いて、テキストプロンプトで提示されたオブジェクトを指定されたオブジェクト位置に一致させるような誘導画像生成を実行しました。

結果の提示 (Presentation of Results) 結果は図 6 に示されています。図 6 から、説明テキスト内のオブジェクトがすべて指定された場所に適切なサイズで配置されていることが確認できます。各場所は、高品質の生成物で満たされており、「ビーチ」から「油絵」まで、多様な画像コンテンツプロンプトに適合しています。

スタイルガイダンス (Style Guidance) 最後に、スタイル画像で与えられる参照スタイルに基づいて画像生成をガイドする実験で、Stable Diffusion を用いた実験を締めくくります。この目的のために、スタイル画像から CLIP の画像特徴抽出器を使用して参照スタイルを取得し、生成画像の埋め込みとスタイル画像の埋め込み間の負のコサイン類似度を損失関数として計算しました。先行実験と同様に、Stable Diffusion モデルにテキスト入力を追加条件付けとして使用してコンテンツを制御しました。

異なるスタイル画像と異なるテキストプロンプトの組み合わせについて実験を行い、結果を図 7 に示しました。図 7 から、生成画像が指定されたテキストプロンプトに一致するコンテンツを含みつつ、与えられたスタイル画像に一致するスタイルを示していることが確認できます。本実験では $s(t) = 6 \cdot \sqrt{1-\alpha_t}$ および $k = 6$ と設定しました。さらに、生成コンテンツの量を制御するため、テキスト条件生成と無条件生成のバランスを取る Stable Diffusion のパラメータ γ を、各列においてそれぞれ 3.0, 3.0, および 4.0 に設定しました。

4.2 4.2 ImageNet Diffusion の結果

本節では、ImageNet で訓練された無条件拡散モデルを用いたガイド付き画像生成の結果を提示します。CLIP ガイダンス、オブジェクト位置ガイダンス、および「セグメンテーションガイド付きイン

ペインティング」と呼ぶハイブリッドガイド付き画像生成タスクの 3 つの手法を検証しました。それぞれのガイダンスに対応する結果と実装を以下で説明します。

CLIP ガイダンス　Stable Diffusion における CLIP ガイダンスと同じ f および ℓ の構成を用いて、CLIP ガイド付き生成を実施しました。本実験ではフォワードガイダンスのみを適用しました。ユニバーサルガイダンスアルゴリズムの限界を評価するため、生成される画像が期待される分布外となるよう手作業でテキストプロンプトを作成しました。具体的には、テキストプロンプトは現実から大きくかけ離れたアートスタイルを指定するか、ImageNet のクラスラベルに該当しないオブジェクトを指定しています。結果を図 8 に示します。結果から、アルゴリズムは高品質の画像生成を成功させ、テキストプロンプトとも一致していることが明らかです。3 つの画像について、それぞれ $s(t) = w \cdot \sqrt{1-\alpha_t}$ と設定し、w は 2、5、2、k は 10、5、10 です。

オブジェクト位置ガイダンス　Stable Diffusion におけるオブジェクト位置ガイダンスと同様に、同じネットワークアーキテクチャと事前学習済みモデルをオブジェクト検出ネットワークとして使用し、ガイダンスアルゴリズム用の同一の損失関数 ℓ を構築しました。ただし、Stable Diffusion とは異なり、オブジェクトの位置情報のみがガイド付き画像生成で使用可能なプロンプトとなります。本実験では、$s(t) = 100\sqrt{1-\alpha_t}$ および $k = 3$ を設定しました。

異なるオブジェクト位置プロンプトを用いて、次の 2 つのアルゴリズム構成を比較しました：(1) フォワードユニバーサルガイダンスのみを使用、(2) フォワードおよびバックワードユニバーサルガイダンスの両方を使用。図 8 から、フォワードとバックワードの両方のガイダンスを適用すると、生成された画像が現実的であり、オブジェクトがプロンプトにうまく一致していることが分かります。一方、フォワードガイダンスのみを使用した場合、画像は現実的であるものの、オブジェクトのカテゴリや位置が一致していないことが確認されました。

5　結果と議論

5.1　セグメンテーションガイド付きインペインティング

この実験では、複数のガイダンス関数を処理するアルゴリズムの能力を探ることを目的としました。インペインティングマスク、分類器、およびセグメンテーションネットワークからの複合ガイダンスを用いたガイド付き画像生成を実施しました。まず、インペインティングのプロンプトとしてマスクされた領域を持つ画像を生成しました。その後、オブジェクトクラス c を分類のプロンプトとして選択し、マスクされた領域を同じクラス c の前景オブジェクトと見なしたセグメンテーションマスクを生成しました。

インペインティングに対しては非マスク領域に対する ℓ_2 損失を損失関数として使用し、対応する $s(t) = 0$、すなわちインペインティングにはバックワードガイダンスのみを使用しました。セグメンテーションネットワークについては、セクション 4.1 で説明したネットワークを使用し、$s(t) = 200\sqrt{1-\alpha_t}$ と設定しました。分類ガイダンスにはノイズのある入力を受け付ける分類器（Dhariwal & Nichol, 2021）を使用し、フォワードガイダンスの代わりに元の分類器ガイダンス（式 (4)）を適用しました。

図 10 に示された結果は、インペインティングと分類器をガイダンスとして使用した場合、アルゴリズムがインペインティングのプロンプトと一致し、指定されたオブジェクトクラスに正しく分類される現実的な画像を生成できることを示しています。さらにセグメンテーションガイダンスを追加すると、生成画像がセグメンテーションマップとインペインティングプロンプトの両方とほぼ完全に一致し、現実感を保つことが確認されました。この結果は、アルゴリズムが個別のガイダンス関数からのフィードバックを効果的に組み合わせて利用できることを示しています。

6 制約

ユニバーサルガイダンスを用いた生成は、標準的な条件付き生成よりも通常遅いという制約があります。経験的に、高品質の画像を生成するには、複雑なガイダンス関数に対して各ノイズレベル t で複数回のデノイジングが必要です。ただし、アルゴリズムの時間計算量は再帰ステップ k の回数に線形にスケールし、k が大きい場合には画像生成が遅くなります。

また、メインペーパーで示されているように、与えられた制約に一致する画像を生成するためには、バックワードガイダンスが必要な場合があります。バックワードガイダンスを計算するには、マルチステップの勾配降下内ループを用いた最適化が必要です。適切な勾配ベースの最適化アルゴリズムと学習率スケジュールの選択により収束速度は大幅に向上しますが、ガイダンス関数自体が非常に大規模なニューラルネットワークである場合、バックワードガイダンスの計算時間は不可避的に長くなります。最後に、最適な結果を得るには、各ガイダンスネットワークに対してサンプリングのハイパーパラメータを個別に選択する必要があることも制約として挙げられます。

7 結論

本論文では、固定された基盤拡散モデルを基に、任意のオフ・ザ・シェルフのガイダンス関数を用いたガイド付き画像生成を実現するユニバーサルガイダンスアルゴリズムを提案しました。このアルゴリズムは、ガイダンス関数や基盤モデルを特定のプロンプトタイプに適応させるための再学習を必要とせず、ガイダンスおよび損失関数が微分可能であれば動作可能です。

本アルゴリズムは、セグメンテーション、顔認識、オブジェクト検出システムなどの複雑なガイダンスを含む実験で有望な結果を示しました。また、複数のガイダンス関数を組み合わせて使用することも可能であることが実証されました。

References

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems, 34:17981–17993, 2021.

Bansal, A., Borgnia, E., Chu, H.-M., Li, J. S., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., and Goldstein, T. Cold diffusion: Inverting arbitrary image transforms without noise. arXiv preprint arXiv:2208.09392, 2022.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. arXiv preprint arXiv:2209.14687, 2022a.

Chung, H., Sim, B., Ryu, D., and Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. arXiv preprint arXiv:2206.00941, 2022b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. IEEE, 2009.

Dhariwal, P. and Nichol, A. Q. Diffusion models beat gans on image synthesis. volume 34, 2021.

Graikos, A., Malkin, N., Jojic, N., and Samaras, D. Diffusion models as plug-and-play priors. arXiv preprint arXiv:2206.09012, 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, 32, 2020.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 1314–1324, 2019.

Huang, R., Lam, M. W., Wang, J., Su, D., Yu, D., Ren, Y., and Zhao, Z. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. arXiv preprint arXiv:2204.09934, 2022.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. arXiv preprint arXiv:2201.11793, 2022.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761, 2020.

Li, X. L., Thickstun, J., Gulrajani, I., Liang, P., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. arXiv preprint arXiv:2205.14217, 2022.

Li, Y., Xie, S., Chen, X., Dollar, P., He, K., and Girshick, R. Benchmarking detection transfer learning with vision transformers. arXiv preprint arXiv:2111.11429, 2021.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11461–11471, 2022.

Neubeck, A. and Van Gool, L. Efficient non-maximum suppression. In 18th International Conference on Pattern Recognition (ICPR ' 06), volume 3, pp. 850–855. IEEE, 2006.

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.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019.

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.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of CVPR, 2022.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. International Conference on Learning Representations, 2021a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations, 2021b.

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, 2022a.

Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490, 2022b.

Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A. G., and Milanfar, P. Deblurring via stochastic refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16293–16303, 2022.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10):1499–1503, 2016.