Factorized Diffusion: Perceptual Illusions by Noise Decomposition

Daniel Geng*, Inbum Park*, and Andrew Owens
University of Michigan
dgeng@umich.edu

Abstract

本論文では、画像を線形成分の和として分解する手法を用い、拡散モデルサンプリングを通じて個々の成分を制御するゼロショット法を提案する。例えば、画像を低空間周波数と高空間周波数に分解し、これらの成分を異なるテキストプロンプトに条件付けることができる。この手法により、観察距離によって外観が変化するハイブリッド画像を生成する。また、画像を 3 つの周波数サブバンドに分解することで、3 つのプロンプトを使用したハイブリッド画像を生成可能である。さらに、画像をグレースケール成分とカラー成分に分解し、暗所照明下で自然に発生する現象である、グレースケールで見たときに外観が変化する画像を生成する。さらに、モーションブラー（動きぼけ）カーネルによる分解を利用し、モーションブラー下で外観が変化する画像を生成する。

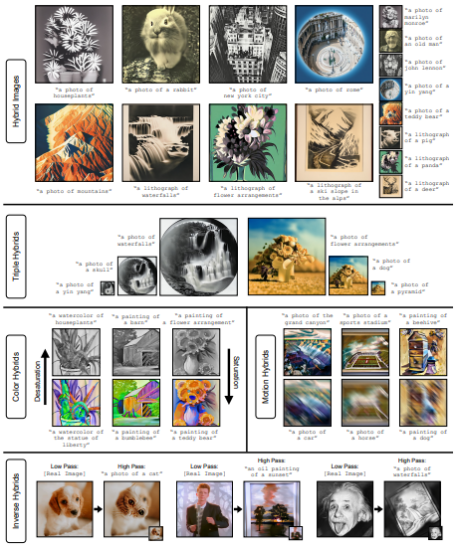
本手法は、異なるプロンプトに条件付けられたノイズ推定の成分から構成される複合ノイズ推定を用いたデノイジングによって動作する。また、特定の分解において、本手法は従来の合成生成および空間制御アプローチを再現することを示す。最後に、本手法を実画像からハイブリッド画像を生成するために拡張できることを示す。これは、1 つの成分を固定し、残りの成分を生成することで実現され、逆問題を効果的に解決する。

キーワード

- 拡散モデル
- 知覚的錯覚
- ハイブリッド画像

1 Introduction

視覚の世界には、画像分解を通じて理解できる現象が数多く存在します。例えば、物体は遠くから見るとぼやけて見え、近くから見ると細部が非常に目立つという、2 つの異なる視点があります。この現象は、周波数空間での分解によって捉えることができます [?, ?]。また、日中はフルカラーで物を見ることができますが、薄



暗い照明下では輝度のみを知覚する現象があり、これは色空間での分解で理解できます。

本研究では、このような分解要素を制御するための簡単な手法を提案します。この手法により、異なる視覚条件下で異なる知覚をもたらしながらも、全体的に一貫性のある画像を生成することが可能です。我々は、このアプローチを用いて様々な知覚的錯覚を生成しました（図 1 参照）。以下は本研究の主要な成果です：

1. **ハイブリッド画像**: Oliva らの古典的研究 [?] に触発され、観察距離によって解釈が変わるハイブリッド画像を生成しました。これを実現するために、生成画像の低周波数成分と高周波数成分を制御しました。さらに、画像を 3 つの周波数サブバンドに分解することで、3 つの異なるプロンプトを用いた「トリプルハイブリッド」を生成しました。

2. **カラー・ハイブリッド**: グレースケールで見たときに外観が変化するカラー画像を生成しました。この現象は薄暗い照明下で自然に発生します。この手法では、画像の輝度を色彩とは別に制御することで実現しました。

3. **モーション・ハイブリッド**: モーションブラー下で外観が変化する画像を生成しました。これは、画像をブラーカーネルを用いて分解することで実現しました。

これらの成果は、異なる条件下で知覚が変化する画像生成を可能にする新しいアプローチを示しています。

2 Our Approach

本研究では、既存の拡散モデルのサンプリング手順に簡単な変更を加えることで、画像生成を制御する新しい手法を提案します。この手法では、画像分解と各成分

を制御するためのテキストプロンプトが与えられた場合、逆拡散プロセスの各ステップでノイズを複数回推定します。各成分に対応するテキストプロンプトに基づき、それぞれのノイズ推定を行い、分解を直接ノイズ推定に適用することで、各成分から合成ノイズ推定を組み立てます（Fig. 2 参照）。特筆すべき点は、本手法がファインチューニング [?, ?] やガイダンスベースの手法 [?, ?, ?, ?, ?] で用いられる補助ネットワークへのアクセスを必要としないことです。

さらに、本手法は既存の画像から成分を取得し、残りの成分をテキスト条件で生成することも可能です。これにより、逆問題を解くための簡単な手法が得られ、既存の拡散モデルを用いた逆問題の解法に関する研究 [?, ?, ?, ?, ?, ?, ?] と深い関連性を持ちます。この技術を実画像からハイブリッド画像を生成するために応用しました。最後に、本手法を特定の分解と組み合わせることで、空間制御 [?] や合成制御 [?] に関する既存の技術を再現できることを示しました。

2.1 Contributions

我々の貢献を以下にまとめます：

- 画像を複数の成分に分解した場合に、これらの成分を画像生成中に制御するための拡散モデルのゼロショット適応手法を提案します。
- 本手法を用いて、以下のような様々な知覚的錯覚を生成しました：
  - 観察距離に応じて外観が変化する画像（ハイブリッド画像）。
  - 照明条件に応じて外観が変化する画像（カラー・ハイブリッド）。
  - モーションブラーが適用されると外観が変化する画像（モーション・ハイブリッド）。

  各錯覚は異なる画像分解に対応しています。
- 従来の手法で生成されたハイブリッド画像と比較し、我々の結果の方が優れていることを示す定量的評価を行いました。
- 本手法がどのように機能し、なぜ効果的であるかを分析し、直感的な理解を提供しました。
- 本手法の単純な拡張により逆問題を解くことができることを示し、このアプローチを実画像からハイブリッド画像を合成するために適用しました。

3 Related Work

3.1 Diffusion Models

拡散モデル [?, ?, ?, ?, ?] は、ガウスノイズが加えられたデータをデノイズするようにトレーニングされます。これは、ノイズのあるデータからノイズを推定し、テキスト埋め込みなどの追加条件を使用することで達成されます。拡散モデルからデータをサンプリングする際には、純粋なガウスノイズを繰り返しデノイズして、最終的にクリーンな画像が得られます。各デノイジングステップは、ノイズ予測

の一部をノイズ画像から除去する更新で構成されます（例: DDPM [?] や DDIM [?]）。拡散モデルの重要な応用例の 1 つは、テキスト条件付き画像生成 [?, ?, ?, ?] であり、本研究ではこれを基盤としています。

3.2 Diffusion Model Control

拡散モデルは、テキストプロンプトに条件付けられて画像を生成および編集する能力を持っています。逆プロセスの修正 [?, ?, ?, ?, ?]、ファインチューニング [?, ?]、テキスト反転 [?, ?, ?, ?, ?]、注意マップの交換 [?, ?, ?]、指示の供給 [?]、またはガイダンスの使用 [?, ?, ?, ?, ?, ?] によって、画像内のコンテンツのスタイル、位置、外観を変更することが比較的容易になっています。

また、合成生成 [?, ?, ?, ?] に関する別の研究では、拡散モデルが複数のテキストプロンプトの組み合わせに従って画像を生成できることが示されています。本研究はこれに基づいており、同様の技術が画像の個々の成分を条件付けることで知覚的錯覚を生成するために適用できることを示します。本研究はまた、Wang ら [?] による研究とも類似しており、拡散モデルを使用してズームビデオを模倣する画像スタックを生成しています。ただし、本研究では、複数の解像度で解釈可能な単一の画像生成に焦点を当てています。

3.3 Computational Optical Illusions

光学的錯覚（オプティカルイリュージョン）は、視覚的に楽しめるだけでなく、人間や機械の知覚を理解する手がかりとしても機能します [?, ?, ?, ?, ?, ?, ?, ?, ?]。そのため、錯覚を生成するための計算手法が多く開発されてきました [?, ?, ?, ?, ?, ?, ?, ?, ?, ?]。

クラシックな研究として、Oliva ら [?] は「ハイブリッド画像」を紹介しました。これは、観察距離や視認時間に応じて外観が変化する画像であり、人間の知覚における多階層処理を利用したものです [?, ?]。具体的には、1 つの画像の低周波数成分と別の画像の高周波数成分を整列させることで、遠くからは前者の画像、近くでは後者の画像として知覚されます。これに対し、本研究の手法は拡散モデルを用いてゼロからハイブリッド画像を生成するため、既存の 2 つの画像を融合させる必要がなく、手動での整列や適切な画像の選定を回避し、アーティファクトも少なくなります。

近年、アーティストや研究者は、テキスト条件付き画像拡散モデルを利用して錯覚を生成しています。例えば、ある匿名アーティストは、QR コード生成モデル [?, ?] を改変してターゲットテンプレートに一致する画像を生成しました [?]。これらも複数の解釈を持つ画像ですが、バイナリマスクテンプレートに限定され、特殊なファインチューニングモデルを必要とします。また、Burgert ら [?] はスコア蒸留サンプリング [?] を利用して、異なる視点から見る、または重ね合わせることで別のプロンプトに一致する画像を生成しました。他の手法としては、Tancik [?] や Geng ら [?] が既存の拡散モデル [?, ?] を使用して、回転、反転、並び替え、歪み、色反転などの変換によって外観が変化する多視点の錯覚画像を生成しています。

これらの手法は、逆拡散プロセス中にノイズのある画像を複数の方法で変換し、それぞれの変換後の画像をデノイズし、ノイズ推定を平均化することで動作

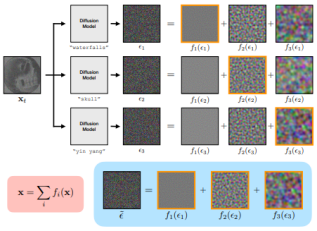


Fig. 2: Factorized Diffusion. Given an image decomposition, we control components of the decomposition through text conditioning during image generation. To do this, we modify the sampling procedure of a pretrained diffusion model. Specifically, at each denoising step, t, we construct a new noise estimate, $\tilde{\epsilon}$, to use for denoising, whose components come from components of ϵ_t, which are noise estimates conditioned on different prompts. Here, we show a decomposition into three frequency subbands, used for creating triple hybrid images, but we consider a number of other decompositions.

します。しかし、ハイブリッド画像で考慮されるような多階層処理のように、ノイズ分布を乱す変換では失敗する場合があります [**?**]。

本研究では、これらのアプローチと同様に、逆拡散プロセスを変更して複数の解釈を持つ画像を生成しますが、ノイズのある画像を操作するのではなく、ノイズ推定を操作する点が異なります。この手法により、従来の研究では扱えなかった錯覚を生成することが可能になります。さらなる議論や結果については、Appendix G を参照してください。

4 Method

画像を複数の成分に分解した場合、本手法ではこれらの成分をテキスト条件付けを用いて制御することが可能です。これを実現するために、テキストから画像を生成する拡散モデルのサンプリング手順を修正します。

4.1 Preliminaries: Diffusion Models

拡散モデルは、ノイズのあるデータを反復的にデノイズすることで分布からサンプリングを行います。T ステップの間、純粋なランダムガウスノイズ x_T をデノイズし、最終ステップでクリーンな画像 x_0 を生成します。中間ステップでは、ノイズのある画像 x_t は以下の形式で表される分散スケジュールに従います：

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

ここで、$\epsilon \sim \mathcal{N}(0, I)$ は標準ガウス分布からのサンプルであり、α_t は事前に設定された分散スケジュールです。

拡散モデルで x_{t-1} を x_t からサンプリングする際、モデル $\epsilon_\theta(\cdot, \cdot, \cdot)$ は時間ステップ t および（オプションで）テキストプロンプト埋め込みのようなコンテキスト y に基づいて、x_t 内のノイズを予測します。その後、更新ステップ $\text{update}(\cdot, \cdot)$ を適用して、ノイズ推定 $\epsilon_\theta := \epsilon_\theta(x_t, y, t)$ の一部をノイズのある画像 x_t から除去します。このステップの具体的な実装は使用する手法に依存しますが、本手法では重要な点として、この更新ステップはしばしば x_t と ϵ_θ（および場合によってはノイズ $z \sim \mathcal{N}(0, I)$）の線形結合として構成されます。

例えば、DDIM [**?**] （$\sigma_t = 0$ の場合）では、更新は次のように実行されます：

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta. \tag{2}$$

4.2 Factorized Diffusion

本手法の概要は Fig. 2 に示されています。本手法では、逆拡散プロセス中にノイズ推定を操作し、推定の異なる成分が異なるプロンプトに条件付けられるようにします。

画像 $x \in R^{3\times H\times W}$ が N 個の成分の和として分解される場合、以下のように表現されます：

$$x = \sum_{i=1}^{N} f_i(x), \tag{3}$$

それぞれの $f_i(x)$ が成分である場合、各成分に異なるテキストプロンプト y_i を対応させることができます。逆拡散プロセスの各ステップで、単一のノイズ推定を計算する代わりに、各 y_i に条件付けられた N 個のノイズ推定を計算します。これを $\epsilon_i = \epsilon_\theta(x_t, y_i, t)$ と表します。その後、各 ϵ_i の成分から構成される合成ノイズ推定 $\tilde{\epsilon}$ を構築します。

$$\tilde{\epsilon} = \sum_i f_i(\epsilon_i), \tag{4}$$

この新しいノイズ推定 $\tilde{\epsilon}$ を使用して、拡散の更新ステップが実行されます。実際には、画像の各成分が異なるテキストプロンプトに条件付けられながらデノイズされ、その結果、異なるプロンプトに基づいて各成分が条件付けられたクリーンな画像が得られます。この手法を「ファクタライズド拡散 (factorized diffusion)」と呼びます。

Sec. 2 で述べたように、本手法は Tancik [**?**] や Geng ら [**?**] の最近の研究に類似しており、視覚的錯覚を生成することを目的としてノイズ推定を修正しています。しかし、本手法は、拡散モデルへの入力 x_t を変更せずにノイズ推定のみを修正する点で異なります。このため、本手法は従来の研究とは異なるクラスの知覚的錯覚を生成します。さらなる議論と結果については、Appendix G をご参照ください。

4.3 3.3 Factorized Diffusion の解析

本手法がなぜ機能するのかを直感的に説明するために、更新関数 update$(\cdot,\cdot)$ がノイズ画像 x_t とノイズ推定 ϵ_θ の線形結合であると仮定します（これは一般的なケースです [**?**, **?**]）。更新関数は t にも依存しますが、簡潔さのためここでは省略します。この場合、更新ステップは以下のように分解できます：

$$x_{t-1} = \text{update}(x_t, \epsilon_\theta), \tag{5}$$

$$= \text{update}\left(\sum_i f_i(x_t), \sum_i f_i(\epsilon_\theta)\right), \tag{6}$$

$$= \sum_i \text{update}\left(f_i(x_t), f_i(\epsilon_\theta)\right), \tag{7}$$

ここで、最初の等式は更新ステップの定義に基づき、2 番目の等式は画像分解を適用することで得られ、3 番目の等式は更新関数の線形性に基づきます。式 (7) は、x_t と ϵ_θ に対する更新ステップが、x_t と ϵ_θ の各成分に対する更新ステップの和として解釈できることを示しています。本手法は、これらの成分それぞれに対して異なる条件付けを使用するものと理解できます。

具体的に書くと、本手法が使用する更新は次のようになります：

$$x_{t-1} = \sum_i \text{update}\left(f_i(x_t), f_i(\epsilon_\theta(x_t, y_i, t)\right). \tag{8}$$

さらに、更新ステップを次のように明示的に記述できます：

$$x_{t-1} = \omega_t x_t + \gamma_t \epsilon_\theta, \tag{9}$$

ここで、ω_t および γ_t は分散スケジュールおよびスケジューラによって決定されます。この場合、f_i が線形であれば以下が成り立ちます：

$$f_i(x_{t-1}) = f_i(\text{update}(x_t, \epsilon_\theta)), \tag{10}$$

$$= f_i(\omega_t x_t + \gamma_t \epsilon_\theta), \tag{11}$$

$$= \omega_t f_i(x_t) + \gamma_t f_i(\epsilon_\theta), \tag{12}$$

$$= \text{update}\left(f_i(x_t), f_i(\epsilon_\theta)\right). \tag{13}$$

これは、x_t の i 番目の成分を ϵ_θ の i 番目の成分で更新すると、x_{t-1} の i 番目の成分にのみ影響を与えることを意味します。

4.4 3.4 考慮した分解方法

本論文で検討した分解方法の詳細を示します。すべての分解に関する結果は Sec. 4 で提示し、議論します。

空間周波数 画像を周波数サブバンドに分解し、それぞれのサブバンドを異なるプロンプトに条件付けすることで、ハイブリッド画像 [**?**] を生成することを目指します。まず、2 つの成分への分解を次のように定義します：

$$f_1(x) = G_\sigma(x), \quad f_2(x) = x - G_\sigma(x), \tag{14}$$

ここで、G_σ は標準偏差 σ を持つガウスぼかしとして実装されたローパスフィルタであり、$x - G_\sigma(x)$ は x のハイパスとして機能します。

さらに、3 つのサブバンドへの分解を考慮し、Fig. 1 で示したトリプルハイブリッド画像を生成します。この場合、成分は次のように定義されるラプラシアンピラミッドのレベルになります：

$$f_1(x) = G_{\sigma_1}(x), \quad f_2(x) = G_{\sigma_2}(x) - G_{\sigma_1}(x), \quad f_3(x) = x - G_{\sigma_2}(x), \tag{15}$$

ここで、σ_1 と σ_2 はおおよそローパス、ミドルパス、およびハイパスのカットオフを定義します。

色空間　色空間による分解も検討し、グレースケールまたはカラーで見たときに異なる解釈を持つ「カラー・ハイブリッド」画像を生成することを目指します。CIELAB 色空間と同様に、画像を明度成分 L と色成分 ab に分解します。CIELAB は知覚的に均一な空間で色を表現することを目指しており、そのため RGB 値の非線形変換を必要とします。これに対して、本研究では単純な線形分解を使用します。

明度成分 L は、画像 x の全ピクセルのチャネル方向平均として定義されます：

$$f_{\text{gray}}(x) = \frac{1}{3}\sum_{c} x_c, \tag{16}$$

ここで、x_c は画像 x のカラー・チャネルを表し、得られる $f_{\text{gray}}(x)$ は x と同じ形状を持ちます。

色成分は以下のように残差として定義されます：

$$f_{\text{color}}(x) = x - f_{\text{gray}}(x). \tag{17}$$

動きぼかし (Motion blurring)　動きぼかしは、ぼかしカーネル K との畳み込みとしてモデル化できます [**?**, **?**, **?**, **?**, **?**, **?**]。動きぼかしが適用されると外観が変化する画像、すなわち「モーション・ハイブリッド (motion hybrids)」を生成するために、次の分解を検討します：

$$f_{\text{motion}}(x) = K * x, \quad f_{\text{residual}}(x) = x - f_{\text{motion}}(x), \tag{18}$$

ここで、画像を動きぼかし成分と残差成分に分割しています。

特に、本研究では一定速度の単純な動きに焦点を当てています。この場合、K はゼロで埋められた行列に、非ゼロ値の線が含まれているものとしてモデル化されます。これはまた、画像を方向性を持つ低周波数成分と残差成分に分解することと同義に考えることもできます。

空間的分解 (Spatial decomposition)　本研究の主な焦点は知覚的錯覚にありますが、分解として空間的マスキングも検討します。画像全体をカバーする互いに素なバイナリ空間マスク m_i が与えられた場合、次の分解を使用できます：

$$f_i(x) = m_i \odot x, \tag{19}$$

ここで、$\odot$ は要素ごとの積を表し、各 $m_i \odot x$ が 1 つの成分となります。この分解の効果は、プロンプトの空間的制御を可能にすることです。これは MultiDiffusion [**?**] の特殊なケースであり、この接続については付録 E で議論します。

スケーリング (Scaling)　最後に興味深い分解方法として、以下の形式を検討します：

$$x = \sum_{i=1}^{N} a_i x, \quad \text{ただし} \quad \sum_{i=1}^{N} a_i = 1\Gamma \tag{20}$$

$a_i = \frac{1}{N}$ とした場合、複数のプロンプトの結合に基づいてサンプルを生成する Liu らの合成拡散法 [**?**] を再現できます。この手法では、ノイズ推定を平均化してサンプリングを行います。

4.5 3.5 逆問題 (Inverse Problems)

生成画像の成分のうち 1 つが既知である場合、たとえば参照画像 x_{ref} から抽出された場合、その成分を固定しながら他の成分を本手法で生成することができます。これにより、実画像からハイブリッド画像を生成することが可能になります (Fig. 1 および Fig. 8 を参照)。

一般性を失うことなく、最初の成分を固定したいと仮定します。この場合、逆拡散プロセスの各ステップの後に x_t を次のように射影します：

$$x_t \leftarrow f_1\left(\sqrt{\bar{\alpha}_t}x_{\text{ref}} + \sqrt{1-\bar{\alpha}_t}\epsilon\right) + \sum_{i=2}^{N} f_i(x_t), \tag{21}$$

ここで、$\epsilon \sim \mathcal{N}(0, I)$、$\alpha_t$ は分散スケジュールによって決定されます。f_1 の引数は、参照画像に対する順方向プロセスのサンプル、つまりタイムステップ t に適したノイズ量を持つ x_{ref} のノイズ付きバージョンです。

基本的に、この手法は x_t を射影し、その最初の成分が x_{ref} の最初の成分と一致するようにします。これは、$y = f_1(x)$ で特徴付けられる（ノイズなしの）逆問題を解くことに相当します。

拡散モデルを事前分布として利用して逆問題を解くための手法は多くの研究で開発されており、本手法のこの拡張は、それらの既存研究の簡略化バージョンと見なすことができます [**?**, **?**, **?**, **?**, **?**]。

5 4 結果

分解方法ごとに整理した結果、逆問題に関する結果、ランダムサンプルの順に結果を提示します。追加の実装詳細は付録 A に、追加の結果は付録 K に記載されています。

5.1 4.1 ハイブリッド画像 (Hybrid Images)

定性的な結果を Fig. 1、Fig. 3、Fig. 4、さらに付録の Fig. 16 に示します。ご覧のように、本手法は高品質なハイブリッド画像を生成します。興味深いことに、ラプラシアンピラミッド分解（式 (15)）を使用することで、3 つの異なるプロンプトを持つハイブリッド画像（トリプルハイブリッド）を生成することも可能でした（Fig. 1 および Fig. 14 参照)。従来の研究 [**?**] では、従来の手法を用いてこれらのトリプルハイブリッドを生成しようとしていましたが、本手法は品質と認識性の点でそれらを大きく上回る結果を示しました（詳細は付録 C を参照)。

ぼかしカーネルの効果 Fig. 3 において、ガウスぼかしの強さ σ が結果にどのように影響するかを示しています。低い σ 値はローパスフィルタのカットオフ周波数が高くなることを意味し、ローパスプロンプトがより目立つ結果となります。σ 値を補間することで、ハイブリッド画像を生成できます。

Oliva ら [?] との比較 Fig. 4 では、Oliva ら [**?**] の手法からのサンプルと本手法によるサンプルを定性的に比較します。[**?**] から直接サンプルを取得し、本手

Fig. 3: Effect of σ. We show a linear sweep over the σ value used in our hybrid decomposition. A lower σ results in the low pass prompt being more prominent, and vice-versa. In between lies hybrid images. ***Best viewed digitally, with zoom.***

Fig. 4: Comparison to Oliva *et al.* [42]. We take hybrid images from Oliva *et al.* [42], and generate our own versions. Left is from our method, and right is from Oliva *et al.*'s. As can be seen, our method produces much more realistic images while still containing both subjects. ***Best viewed digitally, with zoom.***

Table 1: Human Studies. We compare our hybrid images and Oliva *et al.*'s with a two-alternative forced choice test. Participants were shown results from Fig. 4, and were asked which images better contained the prompts, and which were of higher overall quality. Percentages denote the proportion that chose our method. Please see Appendix B for additional details. We find that our method is rated as both higher in quality and better aligned with the prompts. ($N = 77$)

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	Average
High Prompt	70.1%	81.8%	63.6%	51.9%	61.0%	53.2%	70.1%	54.5%	63.3%
Low Prompt	83.1%	84.4%	74.0%	87.0%	93.5%	87.0%	75.3%	81.8%	83.3%
Quality	92.2%	87.0%	83.1%	77.9%	85.7%	79.2%	92.2%	33.8%	78.9%

Table 2: Hybrid Image CLIP Evaluation. We evaluate hybrid images by reporting the maximum clip score over different amounts of blurring. We report the max to compensate for the fact that different hybrid images may be best viewed at different resolutions. Please see Fig. 4 for the referenced hybrid images, and Appendix D for metric implementation details.

	Method	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	Average
Low Pass	Oliva *et al.* [42]	0.268	0.258	0.316	0.250	0.237	0.264	0.257	0.241	0.261
	Ours	0.286	0.252	0.307	0.273	0.275	0.260	0.244	0.269	**0.271**
High Pass	Oliva *et al.* [42]	0.297	0.230	0.306	0.272	0.276	0.306	0.260	0.231	0.272
	Ours	0.321	0.242	0.301	0.258	0.292	0.324	0.320	0.277	**0.292**

法を用いて対応するハイブリッド画像を生成するプロンプトを手動で作成しました。ご覧の通り、本手法で生成されたハイブリッド画像は、望ましいプロンプトを異なる観察距離で保持しつつ、はるかに現実的な結果を示しています。

本手法の利点の 1 つは、低周波数成分と高周波数成分が互いの情報を考慮して生成される点です。これは、拡散モデルが画像全体を入力として受け取るためです。一方で、Oliva らのハイブリッド画像では、周波数成分が 2 つの独立した画像から抽出され、それらを組み合わせる形で生成されます。さらに、これら 2 つの画像は手動で探し出し、整列させる必要がありますが、本手法では単に低周波数成分と高周波数成分が自動的に適切に整列して生成されます。

また、我々のハイブリッド画像と Oliva ら [?] のものを定量的に比較した結果も示します。Tab. 1 では、2 選択強制選択 (2AFC) 調査の結果を示しており、被験者に本手法のハイブリッド画像と Oliva らの画像のどちらを選ぶかを尋ねました。調査では、どちらの画像がプロンプトをより適切に保持しているか、また全体的な品質が高いかを評価しました。詳細については付録 B を参照してください。

結果として、被験者は一貫して、本手法による画像が品質が高く、プロンプトをより適切に反映していると選びました。

CLIP アラインメントスコア　Tab. 2 に CLIP アラインメントスコアを示します。ハイブリッド画像はさまざまな解像度で見るのが最適であるため、プロンプトと異なる程度にぼかした画像との間で最大の CLIP スコアを報告します。メト

Fig. 5: Color Hybrids. We show additional *color hybrid* results. These are images that change appearance when color is added or subtracted away. These images change appearance when moved from bright to dim lighting, in which color is harder to see.

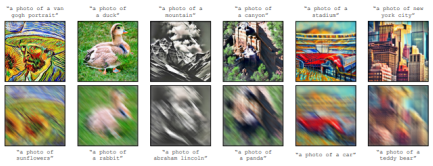


Fig. 6: Motion Hybrids. We show additional *motion hybrid* results. These are images that change appearance when motion blurred. Here, the motion from upper left to bottom right.

リックの実装詳細については付録 D を参照してください。本手法はプロンプトに対してより良いアラインメントを持つハイブリッド画像を生成することがわかります。

5.2 4.2 その他の分解方法

カラー・ハイブリッド 定性的なカラー・ハイブリッドの結果を Fig. 1、Fig. 5、付録の Fig. 17 に示します。ご覧のように、グレースケール画像は 1 つのプロンプトに一致し、カラー画像は別のプロンプトに一致します。例えば、Fig. 5 の「ウサギ」/「火山」の画像では、ウサギの耳がグレースケール画像では溶岩の噴煙として再利用されています。この効果を達成するには、単に任意の量の色をグレースケール画像に追加するだけでは不十分であり、追加された色がグレースケール画像とそのプロンプトとのアラインメントを変化させない必要があります。この技術の興味深い応用例として、明るい照明下と薄暗い照明下で異なる外観を示す画像を生成することが挙げられます。薄暗い環境では、人間の視覚は色を識別するのがはるかに困難です。

モーション・ハイブリッド 定性的なモーション・ハイブリッドの結果を Fig. 1、Fig. 6、付録の Fig. 15 に示します。これらは動きぼかしを加えると外観が変化する画像です。本論文のすべてのモーション・ハイブリッドにおいて、ぼかしカーネル $K = \frac{1}{k} I \in R^{k \times k}$ を使用し、$k = 29$ としています。これは、左上から右下への対角線方向の動きを表しています。

空間的分解 (Spatial decomposition) 画像を互いに素な空間領域に分解し、本手法を適用することで、MultiDiffusion [**?**] の特殊なケースに該当する手法を再現できます。この方法を使用すると、テキストプロンプトが空間的に作用する箇所を細かく制御することが可能です（Fig. 7 参照）。さらなる議論については付録 E を参照してください。

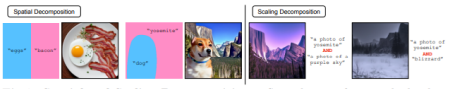


Fig. 7: Spatial and Scaling Decompositions. Special cases of our method reduce to prior work. **(Left)** Decomposing images into spatial regions recovers a special case of MultiDiffusion [2], and allows us to assign prompts to spatial regions. **(Right)** Decomposing an image by scaling allows us to compose concepts, and recovers the method of Liu *et al.* [33].

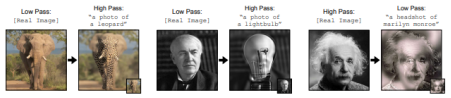


Fig. 8: Hybrids from Real Images. We show hybrid images generated from real images. We take low or high passes of real images, and use our method to fill in the missing component, conditioned on a prompt. ***Best viewed digitally, with zoom.***

Figure 1: Enter Caption

スケーリング分解 (Scaling decomposition)　スケーリング分解で $a_i = \frac{1}{N}$ を使用すると、本手法は Liu ら [?] の拡散モデルにおける構成性に関する先行研究と完全に一致します。具体的には、Liu らが提案した結合演算子を再現します。この結果を Fig. 7 に示しますが、より多くの例については [?] を参照してください。

5.3　4.3 逆問題 (Inverse Problems)

Sec. 3.5 で議論したように、本手法を変更して逆問題を解くことが可能です。この技術は、既存の研究 [?, ?, ?, ?, ?, ?, ?] と非常に類似しています。既存の研究では、拡散モデルを事前分布として使用し、色付け、インペインティング、超解像、位相復元などの問題を解決する方法が探求されています。一方で、本研究ではこのアイデアを実画像からハイブリッド画像を生成する方向に応用します。

具体的には、実画像の低周波成分または高周波成分を取得し、本手法を使用して欠落している成分をプロンプトに基づいて補完します。この結果を Fig. 1 および Fig. 8 に示します。また、色付けの結果を付録 J に示しています。

5.4　4.4 制限事項と負の影響

本手法の主な制限の 1 つは、成功率が比較的低いことです。本手法は安定して適切な画像を生成できますが、非常に高品質な画像はまれです。これについては、ハイブリッド画像、カラー・ハイブリッド、およびモーション・ハイブリッドのランダムサンプルを可視化した Fig. 9 および Fig. 18 で確認できます。この脆弱性は、本手法が拡散モデルにとって大きく分布外の画像を生成することに起因していると考えられます。

さらに、ある成分に関連付けられたプロンプトが他の成分に現れることを防ぐ仕組みがありません。本手法のもう 1 つの失敗ケースとして、1 つの成分に対するプロンプトが生成された画像全体を支配する場合があります。経験的には、プロンプトペアを慎重に選択する（追加の議論については付録 I を参照）か、分解パラメータを手動で調整することで、本手法の成功率を向上させることが可能です。ただし、一般的な手法のロバスト性を向上させることは今後の課題として残します。

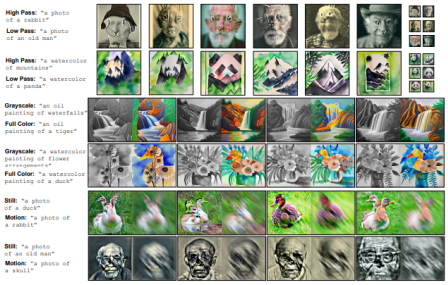


Fig. 9: Random Samples. We provide random samples for selected prompts and decompositions. As can be seen, most random results are of passable quality, with some catastrophic failures, and some very high quality illusions. More random samples are shown in Fig. 18.

強力な画像生成モデルをより良く制御する能力が社会的および倫理的な考慮事項を生むことは明らかです。本手法は錯覚の生成に適用されていますが、ある意味で知覚を欺くことを目的としており、誤情報の拡散といった応用につながる可能性があります。これらの懸念やその他の問題は、さらに研究し慎重に考慮されるべきだと考えています。

6　5 結論

本研究では、拡散モデルのサンプリングを通じて画像の異なる成分を制御するゼロショット手法を提案し、この手法を知覚的錯覚の生成に適用しました。本手法を用いて、ハイブリッド画像、3 つのプロンプトを持つハイブリッド画像、カラー・ハイブリッドやモーション・ハイブリッドといった新しい種類の錯覚を生成しました。また、本手法が動作する理由についての分析と直感的な説明を行いました。

特定の画像分解において、本手法が拡散モデルにおける構成的生成や空間的制御に関する既存研究に一致することを示しました。最後に、逆問題との関連性を明らかにし、この知見を活用して実画像からハイブリッド画像を生成しました。

References

1. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models (2023) 3, 4

2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113* (2023) 3, 4, 8, 12, 13, 20

3. Brooks, T., Barron, J.T.: Learning to synthesize motion blur. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6840–6848 (2019) 8

4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. *CVPR* (2023) 4

5. Burgert, R., Ranasinghe, K., Li, X., Ryoo, M.: Diffusion illusions: Hiding images in plain sight. https://ryanndagreat.github.io/Diffusion-Illusions (Mar 2023) 4, 21, 22

6. Chandra, K., Li, T.M., Tenenbaum, J., Ragan-Kelley, J.: Designing perceptual puzzles by differentiating probabilistic programs. *ACM SIGGRAPH 2022 Conference Proceedings*. pp. 1–9 (2022) 4

7. Chu, H.K., Hsu, W.H., Mitra, N.J., Cohen-Or, D., Wong, T.T., Lee, T.Y.: Camouflage images. *ACM Trans. Graph.* 29(4), 51–1 (2010) 4

8. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687* (2022) 3, 9, 12

9. Chung, H., Sim, B., Ryu, D., Ye, J.C.: Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems* 35, 25683–25696 (2022) 3, 12, 24

10. Chung, H., Sim, B., Ye, J.C.: Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12413–12422 (2022) 3, 12

11. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems* 34, 8780–8794 (2021) 3

12. Du, Y., Li, S., Mordatch, I.: Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems* 33, 6637–6647 (2020) 4

13. Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., Sohl-Dickstein, J.: Adversarial examples that fool both computer vision and time-limited humans. *Advances in Neural Information Processing Systems* 31 (2018) 4

14. Epstein, D., Jabri, A., Poole, B., Efros, A.A., Holynski, A.: Diffusion self-guidance for controllable image generation (2023) 4

15. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. *ACM SIGGRAPH 2006 Papers*, pp. 787–794 (2006) 8

16. Freeman, W.T., Adelson, E.H., Heeger, D.J.: Motion without movement. *ACM SIGGRAPH Computer Graphics* 25(4), 27–30 (1991) 4

17. Geng, D., Owens, A.: Motion guidance: Diffusion-based image editing with differentiable motion estimators. *International Conference on Learning Representations* (2024) 3

18. Geng, D., Park, I., Owens, A.: Visual anagrams: Generating multi-view optical illusions with diffusion models. *Computer Vision and Pattern Recognition (CVPR) 2024* (2024) 4, 5, 6, 21, 22

19. Gomez-Villa, A., Martin, A., Vazquez-Corral, J., Bertalmío, M.: Convolutional neural networks can be deceived by visual illusions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12309–12317 (2019) 4

20. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014) 4

21. Gu, Z., Davis, A.: Filtered-guided diffusion: Fast filter guidance for black-box diffusion models. *arXiv preprint arXiv:2306.17141* (2023) 3, 4

22. Guo, R., Collins, J., de Lima, O., Owens, A.: Ganmouflage: 3D object nondetection with texture fields. *Computer Vision and Pattern Recognition (CVPR)* (2023) 4

23. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022) 4

24. Hertzmann, A.: Visual indeterminacy in GAN art. In: *ACM SIGGRAPH 2020 Art Gallery*, pp. 424–428 (2020) 4

25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239* (2020) 3, 4, 7, 22

26. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly DDPM noise space: Inversion and manipulations. *arXiv preprint arXiv:2304.06140* (2023) 4

27. Jaini, P., Clark, K., Geirhos, R.: Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779* (2023) 4

28. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. *Advances in Neural Information Processing Systems* 35, 23593–23606 (2022) 3, 9, 12, 24

29. Konstantinov, M., Shonenkov, A., Bakshandaeva, D., Ivanova, K.: IF by DeepFloyd Lab at StabilityAI (2023), https://github.com/deep-floyd/IF/, GitHub repository 4, 19

30. Labs, M.: ControlNet QR Code Monster v2 for SD-1.5 (July 2023), https://huggingface.co/monster-labs/control$_v$1p_sd15$_q$r$code_m$onster4$Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion : Coherentmontageviasynchronizedjointdiffusions.In : Thirty-seventhConferenceonNeuralInformationProcessingSystems$(2023)3, 4

31. Liu, N., Li, S., Du, Y., Tenenbaum, J., Torralba, A.: Learning to compose visual relations. *Advances in Neural Information Processing Systems* 34, 23166–23178 (2021) 3

32. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: *European Conference on Computer Vision*. pp. 423–439. Springer (2022) 4, 9, 12, 13

33. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., 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) 3, 9, 12

34. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: *International Conference on Learning Representations* (2022) 4

35. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2502–2510 (2018) 8

36. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models (2022) 4

37. Nayar, S.K., Ben-Ezra, M.: Motion-based motion deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6), 689–698 (2004) 8

38. Ngo, J., Sankaranarayanan, S., Isola, P.: Is clip fooled by optical illusions? (2023) 4

39. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models (2021) 3, 4

40. Oliva, A., Schyns, P.G.: Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology* 34 (1997) 4

41. Oliva, A., Torralba, A., Schyns, P.G.: Hybrid images. *ACM Trans. Graph.* 25(3), 527–532 (Jul 2006). https://doi.org/10.1145/1141911.1141919 1, 2, 4, 7, 10, 11, 20

42. Owens, A., Barnes, C., Flint, A., Singh, H., Freeman, W.: Camouflaging an object from many viewpoints (2014) 4

43. Parmar, G., Singh, K.K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation (2023) 4

44. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2210.02302* (2022) 4, 21

45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), https://github.com/CompVis/latent-diffusion 4

46. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2022) 3, 4

47. 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., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022) 4

48. Schyns, P.G., Oliva, A.: From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science* 5(4), 195–200 (1994) 1, 4

49. Schyns, P.G., Oliva, A.: Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognitive* 69 (1999) 4

50. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F., Blei, D. (eds.) *Proceedings of the 32nd International Conference on Machine Learning*. Proceedings of Machine Learning Research, vol. 37, pp. 2256–2265. PMLR, Lille, France (07–09 Jul 2015). https://proceedings.mlr.press/v37/sohl-dickstein15.html 3

51. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (October 2020), https://arxiv.org/abs/2010.02502 3, 4, 5, 7

52. Song, Y., Shen, L., Xing, L., Ermon, S.: Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005* (2021) 3, 9

53. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: *International Conference on Learning Representations* (2021), https://openreview.net/forum?id=PxTIG12RRHS 3, 12, 24

54. Sripian, P., Yamaguchi, Y.: Hybrid image of three contents. *Visual Computing for Industry, Biomedicine, and Art* 3(1), 1–8 (2020) 9, 20

55. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013) 4

56. Takeda, H., Milanfar, P.: Removing motion blur with space–time processing. *IEEE Transactions on Image Processing* 20(10), 2990–3000 (2011) 8

57. Tancik, M.: Illusion diffusion. https://github.com/tancik/Illusion-Diffusion (Feb 2023) 4, 6

58. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1921–1930 (June 2023) 4

59. Ugleh: Spiral town - different approach to qr monster. https://www.reddit.com/r/StableDiffusion/comme
Edict : $Exactdiffusioninversionviacoupledtransformations.In$: *Proceedings of the IEEE/CVF Confe*

60. Wang, X., Bylinskii, Z., Hertzmann, A., Pepperell, R.: Toward quantifying ambiguities in artistic images. *ACM Transactions on Applied Perception (TAP)* 17(4), 1–10 (2020) 4

61. Wang, X., Kontkanen, J., Curless, B., Seitz, S., Kemelmacher, I., Mildenhall, B., Srinivasan, P., Verbin, D., Holynski, A.: Generative powers of ten. *arXiv preprint arXiv:2312.02149* (2023) 4

62. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490* (2022) 3, 9, 12

63. Wu, C.H., De la Torre, F.: Unifying diffusion models ' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559* (2022) 4

64. Yitzhaky, Y., Mor, I., Lantzman, A., Kopeika, N.S.: Direct method for restoration of motion-blurred images. *JOSA A* 15(6), 1512–1519 (1998) 8

65. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023) 3, 4

66. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. *arXiv preprint arXiv:2303.17076* (2023) 4

A 実装の詳細

A.1 ピクセル拡散モデル

すべての実験で、一般的な潜在拡散モデルではなく、ピクセル拡散モデルである DeepFloyd IF [29] を使用しました。これは、周波数サブバンド、色空間、運動分

解が潜在空間では意味を持たないためです。例えば、潜在空間でチャンネルを平均化しても、解釈可能な画像操作には対応しません。興味深いことに、潜在拡散モデルを使用して潜在コードをぼかすことでハイブリッド画像を構築することは可能ですが、アーティファクトが発生しやすいため（詳細は付録 F 参照）、より一貫性があり理論的なピクセル拡散モデルを選択しました。

A.2 ハイブリッド画像

DeepFloyd IF [29] は、64×64 解像度で画像を生成し、その後 256×256 に拡大します。このため、σ の値は 64×64 スケールに対して指定され、256×256 画像の場合は 4 倍にスケールされます。エッジ効果を最小化するため、両スケールで比較的大きなカーネルサイズ 33 を使用しました。ハイブリッド画像のすべてに対して、$\sigma = 1.0$ から $\sigma = 3.0$ の範囲の値を使用しました。ただし、Fig. 3 の画像については、σ の値を変更して調査しました。

A.3 トリプルハイブリッド

トリプルハイブリッドの合成は非常に難しいため、高品質なサンプルを生成するために σ 値とプロンプトを手動で選択しました。具体的には、Fig. 1 および Fig. 14 に示したトリプルハイブリッドについて、σ_1 は $\sigma_1 = 0.8$ から $\sigma_1 = 1.0$ の範囲、σ_2 は $\sigma_2 = 1.2$ から $\sigma_2 = 2.0$ の範囲で設定しました。

A.4 アップスケーリング

DeepFloyd IF には、256×256 から 1024×1024 へのアップスケールを行う第 3 段階も含まれています。この段階では潜在モデルが使用されるため、本手法は適用していません。アップスケールには、最高周波数成分または色成分に対応するプロンプトのみを使用しました。

B 人間による評価

Amazon Mechanical Turk を用いて人間による評価を実施しました。77 人の「マスターワーカー」に対して、各ハイブリッド画像ペアについて以下の質問を行いました：

- 「どちらの画像が [prompt_1] をより明確に示していますか？」
- 「どちらの画像が [prompt_2] をより明確に示していますか？」
- 「どちらの画像がより高品質ですか？」

低周波数のプロンプトに関する質問では、参加者がコンテンツをより簡単に確認できるように、画像を適宜ダウンサンプリングしました。一方、高周波数のプロンプトに関する質問および品質に関する質問では、画像をフル解像度で表示しました。参加者にはランダムな順序で 8 つのハイブリッド画像ペアを提示しました。

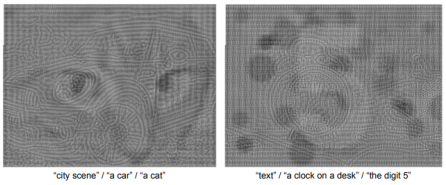


Fig. 10: Prior Work on Triple Hybrid Images. We show the triple hybrid results from prior work [55], which adapts the classic method of [42]. A description of what should be seen is provided underneath each image, going from high to low frequencies. As can be seen, these results are of lower quality than our results.

C 既存のトリプルハイブリッド手法

既存の研究 [55] は、Oliva ら [42] の手法を適応させてトリプルハイブリッド画像を作成しようと試みています。しかし、Fig. 10 に示すように、その結果は高い視覚的品質を持つとは言えず、画像内の 3 つの異なる対象を識別することが困難です。これは、これらの画像を作成する難しさを反映しています。

D メトリクスの実装

Tab. 2 では、複数の画像ダウンサンプリング係数にわたる最大 CLIP スコアを報告しています。具体的には、各ハイブリッド画像について、ダウンサンプリングとアップサンプリングを係数 f で実行し、f を 1 から 8 の間で 20 個の値に線形にスイープしました。これらの画像はその後、CLIP ViT-B/32 モデルの入力解像度である 224×224 サイズに前処理されます。その後、各結果画像埋め込みと対応するプロンプトのテキスト埋め込みの正規化された内積を取り、最大値を報告します。異なるハイブリッド画像が異なるダウンサンプリング係数で最もよく見えることを考慮し、最大値を報告します。

E MultiDiffusion との関連性

Sec. 4.2 では、空間分解を用いた Factorized Diffusion を探求し、プロンプトを特定の空間領域にターゲットすることを可能にすることを示しました。この手法が MultiDiffusion [2] の特殊なケースであると主張します。MultiDiffusion は、複数のノイズ推定値のコンセンサスを画像全体に適用してノイズを除去することで、任意のサイズのノイズ画像を更新します。一方、Factorized Diffusion は空間分解を用いた場合、複数のノイズ推定値の不連続な結合に基づいてコンセンサスを形成します。ただし、私たちの手法は拡散モデルが学習された解像度でのみ動作する点で、MultiDiffusion とは異なります。

F ラテント拡散モデルを用いたハイブリッド画像

Stable Diffusion v1.5 というラテント拡散モデルを用いて本手法を適用した結果を Fig. 11 に示します。この結果から分かるように、生成された画像は一定の品質を持っていますが、ラテント空間でのバンドパスフィルタの適用によって大き

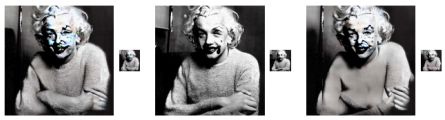


Fig. 11: Latent Hybrid Images. We provide hybrid image results using our method with Stable Diffusion v1.5, a latent diffusion model. As can be seen the results are passable, but suffer from artifacts, due to applying blurring and bandpass operations in the latent space.

なアーティファクトが生じています。ピクセル拡散モデルを使用した場合、はるかに高品質なサンプルが得られることが確認されました。

G 他の手法によるハイブリッド画像の生成

Visual Anagrams [18] と Diffusion Illusions [5] という最近の手法を使用してハイブリッド画像を生成する試みを行いました。その結果を Fig. 12 に示します。両手法とも失敗し、その理由を以下に説明・分析します。

Diffusion Illusions

Diffusion Illusions は、異なるプロンプトとペアリングされた画像の複数ビューに対して SDS [45] ロスを最小化することで動作します。私たちは上述の高パスおよび低パスビューを使用しました。しかし、Fig. 12 に示されているように、この手法は低パスプロンプトの適用には一定の成功を収めましたが、高パスプロンプトを取り込むことに失敗しました。これは、高パスフィルタの適用によって画像が大きく分布外に移動し、SDS の勾配が有効に機能しなくなるためと考えられます。一方で、低パスフィルタは画像の外観を変化させますが、比較的分布内にとどまるため、この手法でも低パスプロンプトの生成が可能でした。

Visual Anagrams

Visual Anagrams は、異なるプロンプトとペアリングされた画像の複数の変換を復元することで動作します。私たちは高パスおよび低パス変換を使用しましたが、これらの操作によってノイズ画像の統計が変化し、手法が失敗しました。その結果、拡散モデルには分布外の画像が供給され、逆過程が収束しないことが Fig. 12 に示されています。

定量評価

最後に、Geng et al. [18] の手法を用いて生成されたハイブリッド画像とカラーのハイブリッド画像を、提案手法と比較しました。その結果を Tab. 3 に示します。プロンプトとして CIFAR-10 の全クラスペアを使用し、各プロンプトペアについて 10 枚の画像をサンプリングし、合計で 900 サンプルを生成しました。評価には [18] と同じメトリクスを使用し、本手法が一貫して良好な結果を示しました。これは、[18] がこれらの錯覚画像を生成するよう設計されていなかったためと考えられます。

Fig. 12: Other Illusion Methods. We attempt to create hybrid images using Visual Anagrams [18] and Diffusion Illusions [5], two recent methods designed to generate optical illusions. As can be seen, both methods fail. Please see Appendix G for analysis.

Table 3: Comparison to Visual Anagrams [18]. We use [18] to synthesize hybrids and color hybrids, and report the same metrics as [18]. We use prompt pairs built from the CIFAR-10 classes, with 10 prompts per pair for a total of 900 samples. Our method performs consistently better, as [18] is not designed to produce these kinds of illusions.

Task	Method	$\mathcal{A}\uparrow$	$\mathcal{A}_{0.9}\uparrow$	$\mathcal{A}_{0.95}\uparrow$	$\mathcal{C}\uparrow$	$\mathcal{C}_{0.9}\uparrow$	$\mathcal{C}_{0.95}\uparrow$
Hybrid Images	Visual Anagrams [18]	0.226	0.237	0.240	0.500	0.520	0.525
	Ours	**0.237**	**0.263**	**0.271**	**0.536**	**0.630**	**0.651**
Color Hybrids	Visual Anagrams [18]	0.223	0.232	0.234	0.500	0.537	0.547
	Ours	**0.231**	**0.260**	**0.269**	**0.512**	**0.562**	**0.586**

H Factorized Diffusion のさらなる解析

Sec. 3.3 で議論したように、私たちの解析では更新ステップがノイズ画像 x_t とノイズ推定 ϵ_θ の線形結合であると仮定しています。しかし、多くの一般的な更新ステップでは、DDPM [25] のようにランダムノイズ $z \sim N(0, I)$ を加えることが含まれます。この場合、更新ステップは以下の 2 つのステップの合成として扱うことができます：

$$x_{t-1} = \text{update}(x_t, \epsilon_\theta), \tag{22}$$

$$= \text{update}(x_t, \epsilon_\theta) + \sigma_z z \tag{23}$$

最初のステップは x_t と ϵ_θ の線形結合であり、2 番目のステップでノイズ z を加えます。したがって、私たちの解析は最初の更新関数 update' に適用されます。

I プロンプトの選択

適切なプロンプトを選択することで、より高品質な錯覚を生成できることが分かりました。例えば、少なくとも 1 つのプロンプトが「柔軟な」被写体（例: 「観葉植物」や「峡谷」）である場合、成功率とサンプルの品質が大幅に向上します。

さらに、分解に特有のバイアスも観察されました。ハイブリッド画像やモーションハイブリッド画像の場合、「写真の...」という形式のプロンプトが良い結果をもたらしました。これは、写真が一般に高周波成分と低周波成分の両方を十分に含むのに対し、「油絵」や「水彩画」のようなスタイルは高周波成分を欠く傾向があるためだと考えられます。

一方、カラーのハイブリッド画像においては、「水彩画」のスタイルを使用することで良い結果が得られました。これは、このスタイルが色彩を強調していることによると推測されます。

J カラー化

Sec. 3.5 で議論したように、逆問題の解決法として私たちの手法を使用し、Fig. 13 でカラー化の結果を示します。具体的には、Sec. 3.4 で導入した色空間の分解を使用します。拡散モデルのサンプリング中に、実画像のグレースケール成分を

Fig. 13: Colorization. Our method can also be used to solve inverse problems, such as colorization. We show grayscale images that we wish to colorize on the left. The color component is then generated conditioned on the text prompts displayed.

固定し、それをカラー化したい画像のグレースケール成分として利用し、カラー成分を生成します。このアプローチは、以前の研究 [9, 28, 54] に非常に類似しています。

K 追加結果

このセクションでは、追加の定性的な結果を提供します。ハイブリッド画像とトリプルハイブリッドの追加結果は、それぞれ Fig. 16 と Fig. 14 に示されています。Fig. 15 と Fig. 17 では、モーションハイブリッドとカラーハイブリッドのさらなる例を提供します。最後に、ハイブリッド画像、カラーハイブリッド、およびモーションハイブリッドのランダムサンプルを Fig. 18 で示します。