Visual Anagrams: Diffusionモデルを用いたマルチビュー錯視画像の生成

Daniel Geng, Inbum Park, Andrew Owens
University of Michigan

Abstract

私たちは、マルチビューの光学錯視を生成する問題に取り組みます。これらは、反転や回転といった変換により外観が変化する画像です。本研究では、市販のテキスト条件付き画像生成用拡散モデルを利用した簡単なゼロショット法を提案します。逆拡散過程において、ノイズの異なるビューを推定し、それらのノイズ推定値を統合して画像を除去します。理論的な解析により、この手法は直交変換として表現できるビュー、特に置換の一部集合として正確に機能することが示唆されます。これに基づき、ピクセルの再配置によって外観が変わる画像「visual anagram」というアイデアが得られます。これには回転や反転、さらにはジグソーパズルのようなピクセル置換といった、よりエキゾチックな変換も含まれます。本手法は、2 つ以上のビューを持つ錯視への自然な拡張も可能です。定性的および定量的な結果により、提案手法の効果と柔軟性を示します。さらなる可視化と結果については、プロジェクトのウェブページをご覧ください: https://dangeng.github.io/visual_anagrams/

1 Introduction

回転や反転といった変換によって外観が変わる画像は、サルバドール・ダリや M. C. エッシャーをはじめとする視覚研究者を長年にわたり魅了してきました。このようなマルチビューの光学錯視の魅力は、視覚要素を複数の異なる方法で理解できるように配置する挑戦にあります。これらの錯視を作成するには、視覚認識を正確にモデル化し、それを巧みに覆す必要があります。

本論文では、市販のテキスト条件付き画像生成用拡散モデルを使用して、マルチビュー錯視を生成するためのシンプルでゼロショットの手法を提案します。従来の光学錯視生成に関する研究の多く [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**] は、人間の視覚認識の明示的なモデルを必要としていましたが、我々の手法はこれを必要としません。むしろ、生成モデルが光学錯視を人間と類似した方法で処理する可能性を示唆する研究 [**?**, **?**, **?**] を基盤としています。この点で、Burgert ら [**?**] や Tancik [**?**] による最近の拡散モデルを用いた光学錯視生成の研究と類似しています。

我々の手法は、画像を反転や回転させることで外観が変わる古典的な錯視（図**??**）や、我々が「ビジュアルアナグラム」と呼ぶ新しいクラスの錯視を生成できます。これらは、ピクセルの置換によって外観が変化する画像です。画像の反転や回転はその一部集合であり、ピクセルの置換として表現可能ですが、さらにエキゾチックな置換も検討します。例えば、2 通りに解けるジグソーパズル「ポリモーフィックジグソー」を生成します。また、3 つまたは 4 つのビューを持つ錯視を生成することにも成功しました（図**??**）。

本手法は、拡散モデルを用いて複数のビューからノイズを除去し、複数のノイズ推定値を得ることで機能します。これらのノイズ推定値を組み合わせ、逆拡散過程のステップに使用する単一のノイズ推定値を形成します。ただし、これらのビューを選択する際には注意が必要です。特に、変換はノイズの統計量を保持する必要があります。拡散モデルは独立同分布（i.i.d.）のガウスノイズを前提として訓練されているためです。この条件の分析を行い、本手法がサポートする変換クラスの正確な仕様を示します。

本研究の貢献は以下の通りです：

- 拡散モデルを用いたマルチビュー光学錯視を生成するためのシンプルで効果的な手法を提案します。
- 本手法がサポートするビューの集合を正確に記述し、それらが機能することを実証的に示します。
- 生成する錯視の品質を最適化するための実用的な設計決定を検討し、それに関するアブレーションを報告します。
- 提案手法の効果と柔軟性を示す定量的および定性的な結果を提示します。

2 Related Work

2.1 Diffusion Models

拡散モデル [**?**, **?**, **?**, **?**, **?**] は、ノイズ分布からデータ分布へのサンプルを反復的に変換する強力な生成モデルの一種です。このモデルは、ノイズのあるサンプル内のノイズを推定し、DDPM [**?**] や DDIM [**?**] などの更新ルールに従っ

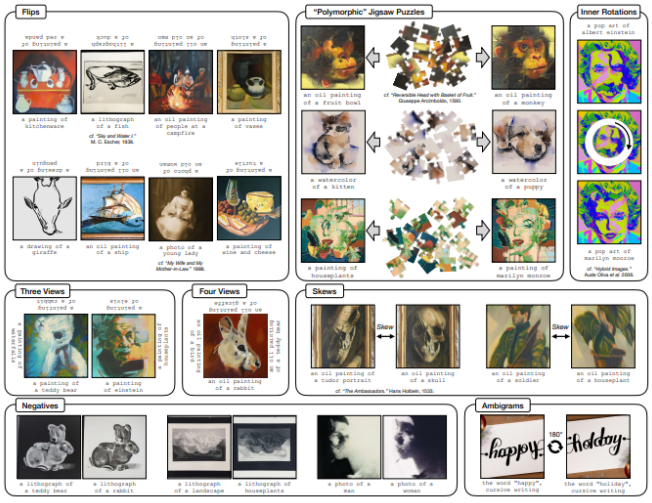

Figure 1. **Generating Multi-View Illusions.** We propose a method for generating optical illusions from an off-the-shelf text-to-image diffusion model. We create images that match different prompts after undergoing a transformation. Our approach supports a variety of transformations, including flips, rotations, skews, color inversions, and jigsaw rearrangements. All images are hand selected. For random samples, please see Fig. 8 and Appendix D. **For easier viewing, please see our webpage for animated versions of these illusions.**

て推定されたノイズを除去することで機能します。拡散モデルの顕著な応用例として、テキスト条件付きの画像生成 [?, ?, ?, ?] が挙げられます。これらのモデルは、ノイズのある画像やタイムステップに加えて、テキストプロンプトの条件として言語モデルの埋め込みを使用します。我々のアプローチは、エネルギーベースモデルと拡散モデルの構成を試みる最近の研究 [?, ?, ?, ?] と密接に関連しています。これらのアプローチ [?, ?] は、複数の条件付き分布から得られるノイズ推定値を組み合わせることで、学習した分布の組み合わせからサンプルを取得できることを示しています。我々の手法も同様のアプローチを使用し、マルチビュー錯視生成の問題に適用しています。

2.2 Computational Optical Illusions

光学錯視は、人間および機械の知覚を理解するためのテストベッドとして機能します [?, ?, ?, ?, ?]。本研究では、錯視を計算的に生成することに焦点を当てています。この分野は主に、外部刺激を処理する人間の脳のモデルに依存してきました。Freeman ら [?] は、局所的なフィルタに連続的にシフトする位相を適用することで、特定の方向への一定の運動の錯覚を作成しました。この方法は、局所的な位相シフトが全体的な運動として解釈されるという観察に基づいています。Oliva ら [?] は、「ハイブリッド画像」を作成する手法を提案しました。この画像は、見る距離に応じて外観が変化します。この方法は、人間の知覚が持つマルチスケールの特性を利用し、1 つの画像の高周波成分と別の画像の低周波成分をブレンドします。Chu ら [?] は、オブジェクトのテクスチャを再構成し、さらに輝度の制約を加えることで、シーン内のオブジェクトをカモフラージュしました。また、他の研究では、3D シーン内の複数の視点からオブジェクトをカモフラージュする手法も提案されています [?, ?]。

最近では、Chandra ら [?] が、人間の視覚をベイズモデルとして微分処理することで、色の恒常性やサイズの恒常性、顔の知覚錯視を設計しました。我々の手法も錯視を生成しますが、人間の知覚の明示的なモデルには依存していません。その代わりに、データを通じて暗黙的に学習された拡散モデル内の視覚的事前情報を活用します。この点は、生成モデルが人間と同様に錯視を処理し、同じ曖昧性を予測するという観察 [?, ?, ?] と一致しています。この観点から、我々の手法は、識別モデルではなく生成モデルを活用して、人間に対する敵対的例 [?] を合成するものと見なすことができます [?]。

2.3 Illusions with Diffusion Models

非常に最近、アーティストや研究者たちは拡散モデルを利用して錯視を作成する可能性を示す取り組みを始めています。偽名で活動するアーティスト MrUgleh [?] は、QR コード生成用に微調整されたモデル [?, ?] を再利用し、指定されたテンプレート画像に微妙に一致するグローバルな構造を持つ画像を作成しました。それに対して、我々は既存の拡散モデルを用いたゼロショットの方法で作成可能なマルチビュー錯視を研究しています。我々の錯視は画像ではなくテキストを通じて指定されます。

Burgert ら [?] はスコア蒸留サンプリング（SDS）[?, ?] を使用して、異なるビューから異なるプロンプトに一致する画像を作成しました。このアプローチは理論上、我々のビューのスーパーセットをサポートしますが、SDS の使用により著しく低品質な結果となり、明示的な最適化が必要となるため、サンプリング時間が長くなるという問題があります。

我々の方法は Tancik [?] による概念実証に最も類似しています。Tancik は、潜在拡散モデル [?] を用い、異なるビューとプロンプト間でノイズ推定を交互に行いながらサンプリングすることで、回転錯視を作成しました。我々の

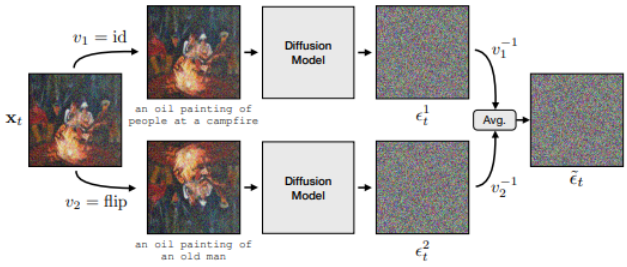

Figure 2. **Algorithm Overview.** Our method works by simultaneously denoising multiple views of an image. Given a noisy image $\mathbf{x}_t$, we compute noise estimates, ϵ_t^i, conditioned on different prompts, after applying views v_i. We then apply the inverse view v_i^{-1} to align estimates, average the estimates, and perform a reverse diffusion step. The final output is an optical illusion.

技術的アプローチは類似していますが、対照的に、我々はマルチビュー錯視を体系的に研究しています。具体的には、多様な種類の錯視を実験的に評価するだけでなく、どのビューがサポートされ、どのビューがサポートされないかについての理論的分析を提供しています。

これにより、単なる回転ビューを超えた成果を得ることができました。また、潜在拡散から発生するアーティファクトの原因を特定し、任意の数のビューをサポートする機能を追加するなど、質的および量的に優れた錯視を作成するためのいくつかの改善を行いました。これらのアプローチによって生成された錯視を体系的に評価したのは、我々が初めてであると考えています。

3 Method

我々の目標は、事前学習済みの拡散モデルを使用してマルチビュー光学錯視を生成することです。すなわち、回転や反転などの変換時に外観やアイデンティティが変化する画像を生成しようとしています。

3.1 Text-conditioned Diffusion Models

拡散モデル [?, ?, ?] は、独立同分布のガウスノイズ x_T を受け取り、それを反復的に除去して、あるデータ分布からサンプル x_0 を生成します。これらのモデルは、部分的にノイズが除去された中間データポイント x_t のノイズを推定するニューラルネットワークによってパラメータ化されており、その推定値は $\epsilon_\theta(x_t, y, t)$ と表されます。ここで、y はテキストプロンプトなどの条件付けであり、t は拡散プロセスのタイムステップを表します。推定されたノイズは更新則 [?, ?] に使用され、x_t から x_{t-1} を計算します。

拡散モデルを別の入力（例えばテキストプロンプト）に条件付けするため、一般的なアプローチとして **classifier-free guidance** [?] が用いられます。この方法では、無条件のノイズ推定（通常、条件付けとして空のテキストプロンプトを渡すことで得られるもの）と条件付きノイズ推定を組み合わせます：

$$\epsilon_t^{\text{CFG}} = \epsilon_\theta(x_t, t, \emptyset) + \gamma\left(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t, \emptyset)\right). \tag{1}$$

ここで、$\emptyset$ は空文字列の埋め込みを示し、γ はガイダンスの強度を制御するパラメータです。

Classifier-free guidance は、生成される画像の分布をシャープにし、高品質な結果を生み出すために機能します。また、**negative prompting** [?] を可能にします。この手法では、空のテキストプロンプト埋め込み $\emptyset$ を、モデルが生成しないようにしたいテキストプロンプトに置き換えます。

3.2 Parallel Denoising

我々は、拡散モデルを用いて画像の複数のビューを同時にデノイズすることにより、マルチビュー錯視を生成します。具体的には、N 個のプロンプト y_i と、それぞれが画像に変換を適用するビュー関数 $v_i(\cdot)$ のセットを取ります。これらの変換は、例えば恒等関数、画像の反転、またはピクセルの置換などが含まれます。そして、拡散モデル $\epsilon_\theta(\cdot)$ と部分的にデノイズされた画像 x_t を用いて、異なるビューからのノイズ推定を平均化して単一のノイズ推定を生成します：

$$\tilde{\epsilon}_t = \frac{1}{N}\sum_i v_i^{-1}\left(\epsilon_\theta(v_i(x_t), y_i, t)\right). \tag{2}$$

本手法では、各ビュー v_i を使用してノイズのある画像 x_t を変換し、変換された画像のノイズを推定し、その後、推定値に v_i^{-1} を適用して元のビューに戻します。これらのノイズ推定値の平均を取ることで、結合されたノイズ推定値を得ることができます。この結合ノイズ推定値は、選択した拡散サンプラーとともに使用されます。

このノイズ推定を結合する手法は、構成可能性に関する以前の研究 [?, ?, ?, ?, ?, ?] と類似しており、その詳細はこれらの研究で議論されています。Classifier-free guidance を組み込むためには、推定値 $\epsilon_\theta(v_i(x_t), y_i, t)$ をその classifier-free 推定値 ϵ_t^{CFG} に置き換えるだけです。

3.3 Conditions on Views

ビューに関する条件の一つとして、まず考えられるのは、それらが可逆である必要があるということです。しかし、拡散モデル ϵ_θ はビュー $v_i(\cdot)$ に対して暗黙的に他の条件も課しています。以下にそのような条件を二つ説明します。これらの条件が満たされない場合、デノイズプロセスは不良な結果を生じることがわかっています。

Linearity（線形性） 拡散モデル ϵ_θ はノイズ画像 x_t に対して動作します。具体的には、次の形の画像を扱います：

$$x_t = w_t^{\text{signal}} x_0 + w_t^{\text{noise}} \epsilon, \tag{3}$$

ここで、w_t^{signal} および w_t^{noise} の正確な値は、分散スケジュールのようなモデルの実装詳細に依存しますが、本研究では重要ではないため、簡潔さを保つために省略します。重要なのは、x_t が純粋な信号 x_0 と純粋なノイズ ϵ の線形結合であることです。したがって、ビュー v_i はノイズ画像 x_t を取り、それを純粋な信号と純粋なノイズの同じ重み付けを持つ新しいノイズ画像 $v_i(x_t)$ に変換する必要があります。

この条件を満たすには、v_i が以下の形式の線形変換である必要があります：

$$v_i(x_t) = A_i x_t, \tag{4}$$

ここで、A_i はある行列であり、x_t はフラット化されたノイズ画像です。

線形性により、ビュー v_i を信号とノイズに個別に適用することが可能になります：

$$v_i(x_t) = A_i(w_t^{\text{signal}} x_0 + w_t^{\text{noise}} \epsilon) \tag{5}$$

$$= w_t^{\text{signal}} A_i x_0 + w_t^{\text{noise}} A_i \epsilon. \tag{6}$$

この結果として、変換された信号 $A_i x_0$ と変換されたノイズ $A_i \epsilon$ の線形結合が得られ、それぞれが適切なスケーリング係数で重み付けされています。さらなる議論については、付録 H を参照してください。

Statistical Consistency（統計的一貫性） 線形結合における信号とノイズの特定の重み付けが期待されるだけでなく、拡散モデルはノイズが特定の分布を持つことも期待します。特に、多くの拡散ネットワークは $\epsilon \sim \mathcal{N}(0, I)$ の条件で訓練されています。したがって、変換されたノイズ $A_i \epsilon$ が同様に $\mathcal{N}(0, I)$ に従うことを保証する必要があります。

この条件が満たされるのは、A_i が直交行列である場合に限られます。この事実の証明は付録 I に記載していますが、直感的には、この事実は標準ガウス分布の球対称性を反映しています。直交変換は、高次元空間における回転や反転の一般化であり、この球対称な分布を保持します。

なお、ここでの回転は画素値の回転を指しており、空間的な回転とは異なります。

3.4 3.4 Views Considered（考慮された変換）

直交変換の大半は、直感的な画像変換には対応しません。しかし、それらの中には視覚的に意味のある変換も存在します。本節では、考慮された直交変換を列挙します。これらの変換は、特に明記されていない限り、図 1 の錯視に含まれています。

Identity（恒等変換） 最も単純な変換は恒等変換です。このビューを用いることで、選択されたプロンプトに一致するように未変換の画像を最適化することが可能です。

Standard Image Manipulations（標準的な画像操作） 画像の空間的回転も考慮します。これは画素の置換として解釈できます。この置換は直交変換の一種であるため有効です。ただし、回転ビューを適用する際には注意が必要です。例えば、バイリニア補間のような一般的なアンチエイリアス処理はノイズの統計特性を変化させます。この点については 4.4 節でさらに議論します。また、空間的な反射も画素の置換であるため、このビューを使用して錯視を生成できます。最後に、列の画素を異なる変位でロールさせることで、スキュー（傾き）の近似も実装しました。

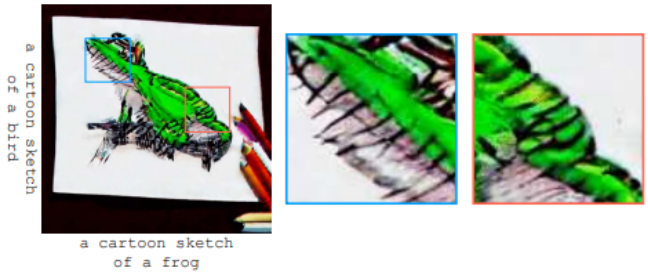

Figure 3. **Latent-Based Artifacts.** Manipulating the *location* of latent codes does not change the *orientation* of the blocks for which they encode. Therefore, when using latent diffusion models we see artifacts as shown above, in which straight lines are thatched under a rotation.

General Permutations（一般的な置換） 空間的な回転、反射、スキューといった特殊ケースを考慮した上で、他の置換も検討します。例えば、画像をジグソーピースに分割し、それらを再配置することで、2 つの解を持つジグソーパズルを生成できます。これを「ポリモーフィックジグソーパズル」と呼びます。実装の詳細は付録 F に記載しています。

また、画素の完全にランダムな置換をサンプリングし、それをビューとして扱う極端なケースも考慮します。この複雑さを軽減するため、画素ではなく正方形のパッチの置換を用いる場合もあります。これらの錯視の例は図 6 に示され、4.3 節で議論されています。

最後に、画像内の円を回転させ、それ以外の部分を静止させる「内回転」も考慮しました。ここで検討した置換は包括的なものではなく、さらに多くの巧妙な変換が存在する可能性がありますが、それらは本研究では扱っていません。

3.5 Color Inversion（色反転）

ネゲーション（色反転）は直交変換の一種であり、直感的には高次元で一般化された 180 度回転に相当します。この特性を利用して、色が反転することで見た目が変わる錯視を生成することができます。なお、この操作が機能するためには、画素値が 0 を中心に配置されている必要があります（例えば、範囲が $[-1, 1]$ の場合）。

3.6 Arbitrary Orthogonal Transformations（任意の直交変換）

ピクセル空間で画像を任意に回転させる場合、その結果は直感的には解釈不能となります。それにもかかわらず、我々の手法がこれらの変換にも適用可能であることを示しました。これらの「錯視」は人間の目には解釈できないものですが、任意の直交変換が本手法においてビューとして機能することの確認になります。これらの例は図 7 に示され、4.3 節で議論されています。

3.7 3.5 Design Decisions（設計上の決定事項）

基本的な手法に加え、錯視の品質を最大化するための設計上の選択肢について検討しました。

Pixel Diffusion Model（ピクセルベースの拡散モデル） 先行研究 [42] では、Stable Diffusion[36] という潜在拡散モデルを用いて多視点のデノイジングが行われました。しかし、潜在表現は実質的に画素のパッチをエンコードするため、回転や反転の際にアーティファクトが発生します。この場合、潜在表現の位置は変化しますが、そのブロックの内容や向きは変わりません。図 3 では、このモデルが 90 度回転下で直線を生成するために格子状の線を描くことを余儀なくされる例を示しています。

この問題を軽減するために、我々はピクセルベースの拡散モデルである DeepFloyd IF[24] を用いて手法を実装しました。DeepFloyd はピクセル上で直接デノイジングを行うため、潜在コードブロックの向きの問題を回避できます。

Combining Noise Estimates（ノイズ推定値の結合） 異なるビューからのノイズ推定値の平均を取るだけでなく、タイムステップごとにそれらを交互に使用する方法も検討しました。この場合、以下のようにして推定値を計算します：

$$\tilde{\epsilon}_t = v_t^{-1}{}_{\bmod N}\left(\epsilon_\theta\left(v_{t \bmod N}(x_t), t, y\right)\right). \tag{7}$$

これは [42] で使用された手法ですが、4.2 節で示すアブレーション研究では、平均を取る方法に比べて性能が劣ることがわかりました。

Table 1. **Quantitative Results.** We report the alignment score, $\mathcal{A}$, and the concealment score, $\mathcal{C}$, as well as quantiles of these scores. For a discussion, please see Sec. 4.1.

Prompt Pair	Method	$\mathcal{A}\uparrow$	$\mathcal{A}_{0.9}\uparrow$	$\mathcal{A}_{0.95}\uparrow$	$\mathcal{C}\uparrow$	$\mathcal{C}_{0.9}\uparrow$	$\mathcal{C}_{0.95}\uparrow$
CIFAR	Burgert *et al.* [2]	0.225	0.253	0.260	0.501	0.526	0.537
	Tancik [42]	0.278	0.310	0.316	0.595	0.692	0.712
	Ours	**0.287**	**0.321**	**0.327**	**0.624**	**0.717**	**0.739**
Ours	Burgert *et al.* [2]	0.233	0.270	0.283	0.501	0.526	0.538
	Tancik [42]	0.256	0.294	0.309	0.545	0.621	0.655
	Ours	**0.275**	**0.315**	**0.326**	**0.574**	**0.668**	**0.694**

Table 2. **Ablations.** We ablate negative prompting, reduction methods, and guidance scales on our dataset.

Ablation	$\mathcal{A}\uparrow$	$\mathcal{A}_{0.9}\uparrow$	$\mathcal{A}_{0.95}\uparrow$	$\mathcal{C}\uparrow$	$\mathcal{C}_{0.9}\uparrow$	$\mathcal{C}_{0.95}\uparrow$
Negative Prompting	0.24	0.27	0.276	**0.576**	**0.659**	**0.683**
No Negative Prompting	**0.255**	**0.285**	**0.295**	0.567	0.643	0.679
Alternating Reduction	0.252	**0.286**	0.292	0.560	0.639	0.664
Mean Reduction	**0.255**	0.285	**0.295**	**0.567**	**0.643**	**0.679**
$\gamma = 3.0$	0.239	0.271	0.285	0.537	0.610	0.629
$\gamma = 7.0$	0.255	0.285	0.295	0.567	0.643	0.679
$\gamma = 10.0$	**0.259**	**0.290**	**0.297**	**0.576**	**0.664**	**0.702**

Negative Prompting（ネガティブプロンプト） 2 ビューのケースで、片方のビューのプロンプトをもう片方のビューのネガティブプロンプトとして使用する実験を行いました。これにより、特定のビューに対して他のビューのプロンプトを隠すことをモデルに促します。詳細な議論は 4.2 節のアブレーション研究をご覧ください。

4 4. Results（結果）

本節では、定量的および定性的な結果、さらに定量的アブレーションを示します。特に指定がない場合、定性的な結果は品質を基準に手動で選定されたものです。ランダムサンプルについては図 8 および付録 D を参照してください。すべての実装詳細は付録 A に記載されています。

4.1 4.1 Quantitative Results（定量的結果）

Metrics（評価指標） 生成されたビューが期待されるプロンプトとどの程度一致するかを測定するために、CLIP[34] を使用しました。具体的には、スコア行列 $S \in R^{N\times N}$ を次式で定義します：

$$S_{ij} = \phi_{\text{img}}(v_i(x))^T \phi_{\text{text}}(p_j), \tag{8}$$

ここで、ϕ_{img} および ϕ_{text} はそれぞれ CLIP の視覚およびテキストエンコーダを表し、単位ノルムのベクトル埋め込みを返します。x は生成された錯視画像、v_i は対応するプロンプト p_i と結びついたビューを表します。ドット積が高いほど、画像とテキストの類似度が高いことを示します。

最初の評価指標として、min diag(S) を考慮します。この指標は、すべてのビューの中で最も低い整合性を直感的に測定します。この指標を **A**（Alignment Score、整合スコア）と呼びます。しかし、この指標はビュー v_j においてプロンプト p_i（$i \neq j$）が見える可能性を考慮していません。このようなケースは本手法の時折見られる失敗例であり、これを定量化するために 2 つ目の指標を提案します。

この 2 つ目の指標を **C**（Concealment Score、隠蔽スコア）と呼び、次式で計算します：

$$C = \frac{1}{N}\text{tr}(\text{softmax}(S/\tau)), \tag{9}$$

ここで、τ は CLIP の温度パラメータです。この指標を計算する際には、softmax の両方向を平均化し、この指標がビューが N 個のプロンプトのいずれかに適切に分類される程度を測定するようにします。

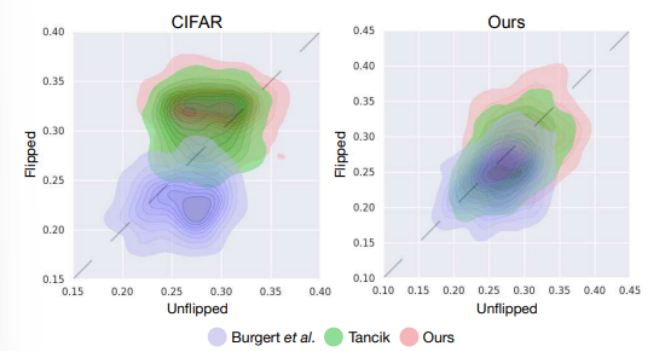

Figure 4. **Flip View CLIP Score Distribution.** We visualize trade-offs between flipped and unflipped views by plotting the distribution of CLIP scores on the datasets. Note that the quality of the flipped image is as good as the unflipped image, with parity indicated by the dashed line.

5 4. Results（結果）

本節では、定量的および定性的な結果、さらに定量的アブレーションを示します。特に指定がない場合、定性的な結果は品質を基準に手動で選定されたものです。ランダムサンプルについては図 8 および付録 D を参照してください。すべての実装詳細は付録 A に記載されています。

5.1 4.1 Quantitative Results（定量的結果）

Metrics（評価指標） 生成されたビューが期待されるプロンプトとどの程度一致するかを測定するために、CLIP[34] を使用しました。具体的には、スコア行列 $S \in R^{N \times N}$ を次式で定義します：

$$S_{ij} = \phi_{\text{img}}(v_i(x))^T \phi_{\text{text}}(p_j), \tag{10}$$

ここで、ϕ_{img} および ϕ_{text} はそれぞれ CLIP の視覚およびテキストエンコーダを表し、単位ノルムのベクトル埋め込みを返します。x は生成された錯視画像、v_i は対応するプロンプト p_i と結びついたビューを表します。ドット積が高いほど、画像とテキストの類似度が高いことを示します。

最初の評価指標として、$\min \text{diag}(S)$ を考慮します。この指標は、すべてのビューの中で最も低い整合性を直感的に測定します。この指標を **A**（Alignment Score、整合スコア）と呼びます。しかし、この指標はビュー v_j においてプロンプト p_i（$i \neq j$）が見える可能性を考慮していません。このようなケースは本手法の時折見られる失敗例であり、これを定量化するために 2 つ目の指標を提案します。

この 2 つ目の指標を **C**（Concealment Score、隠蔽スコア）と呼び、次式で計算します：

$$C = \frac{1}{N} \text{tr}(\text{softmax}(S/\tau)), \tag{11}$$

ここで、τ は CLIP の温度パラメータです。この指標を計算する際には、softmax の両方向を平均化し、この指標がビューが N 個のプロンプトのいずれかに適切に分類される程度を測定するようにします。

5.2 Dataset and Baselines（データセットとベースライン）

Dataset（データセット） 本手法およびベースラインの評価を行うため、2 ビュー錯視のプロンプトペアからなる 2 つのデータセットを構築しました。1 つ目のデータセットは CIFAR-10 の 10 クラスを使用し、各クラスペアに 1 つのプロンプトを割り当てた合計 45 のプロンプトペアを含みます。このデータセットを「CIFAR」と呼びます。2 つ目のデータセットは手作業で構築したもので、そのプロセスは付録 B に記載されています。このデータセットには 50 のプロンプトペアが含まれ、「Ours」と呼びます。

Baselines（ベースライン） 本研究では、既存の拡散モデルを使用して錯視を生成する 2 つのベースラインを使用します。1 つ目は「Burgert et al. [2]」とし、Score Distillation Sampling（SDS）を用いたものです。2 つ目は「Tancik [42]」とし、本手法の初期バージョンであり、詳細な違いについてはセクション 2 で議論されています。

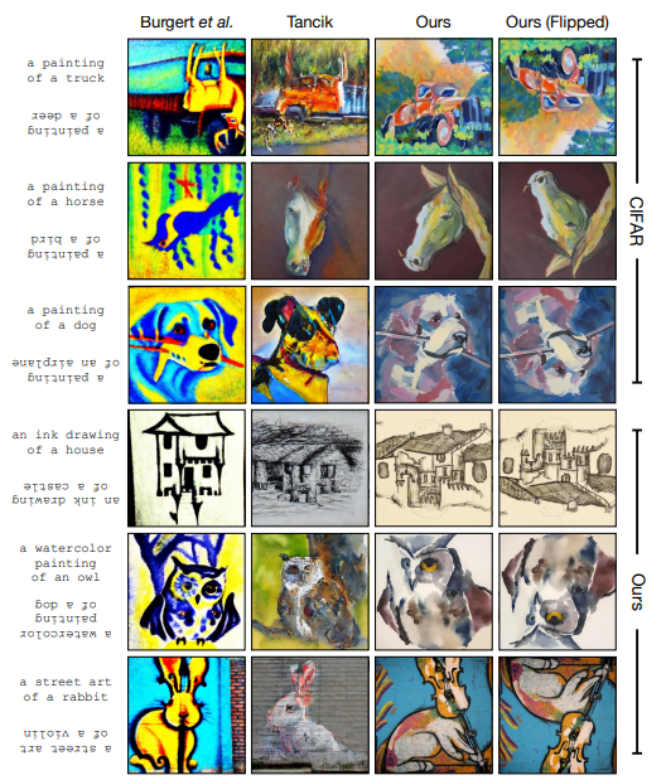

Figure 5. **Qualitative Comparisons.** We compare illusions generated by baselines to our illusions. We show examples from both our prompt dataset and the CIFAR prompt dataset.

Results（結果） 表 1 では、本手法とベースラインを両データセット上で比較した結果を示します。垂直反転を使用しましたが、これは本手法およびベースラインの両方でサポートされている変換です。各プロンプトに対して 10 サンプルを生成し、CIFAR データセットでは合計 450 サンプル、Ours データセットでは 500 サンプルを生成しました。Burgert et al. の手法は SDS を使用しており、非常に遅い 1 ため、より多くのサンプルを使用して公平な比較を行うことは困難です。

特に「ベストケース」の性能に注目するため、メトリックの分位点も報告します。たとえば 90 パーセンタイルを示すものを $A_{0.9}$ とします。結果として、本手法は整合スコア（Alignment Score）および隠蔽スコア（Concealment Score）の両方で一貫してベースラインを上回っています。

また、2 つのビューを最適化する際のトレードオフをより明確に理解するため、錯視の 2 つのビューそれぞれにおける CLIP スコアをプロットした密度プロットを図 4 に示します。この結果から、本手法は平均およびベストケースの両方でベースラインよりも優れています。さらに、デノイジング中に反転を行っても性能に影響を与えません。反転した画像の品質は非反転画像と同等に高い結果を示しています。

5.3 4.3 定性的結果

本節では、図 1、図 5、図 6、および図 7 に示す定性的な結果を示します。ランダムサンプルについては、図 8 および付録 D を参照してください。また、追加の定性的サンプルは付録 C に示しています。全体的に、本手法は幅広いビューに対して非常に高品質な視覚的錯視を生成できることが分かりました。興味深いことに、本手法は、1 つのビューの要素を他のビューに再利用する巧妙な方法を見つけることがよくあります。例えば、図 1 に示す「滝」/「ウサギ」/「テディベア」の 3 ビュー錯視では、テディベアの鼻がウサギの目であり、滝の中の岩であるといった例があります。

Baselines（ベースライン） 本手法とベースライン手法の定性的な比較を図 5 に示します。各手法について 100 サンプルの中から最良の画像を選択しました。結果として、本手法が優れていることが分かります。

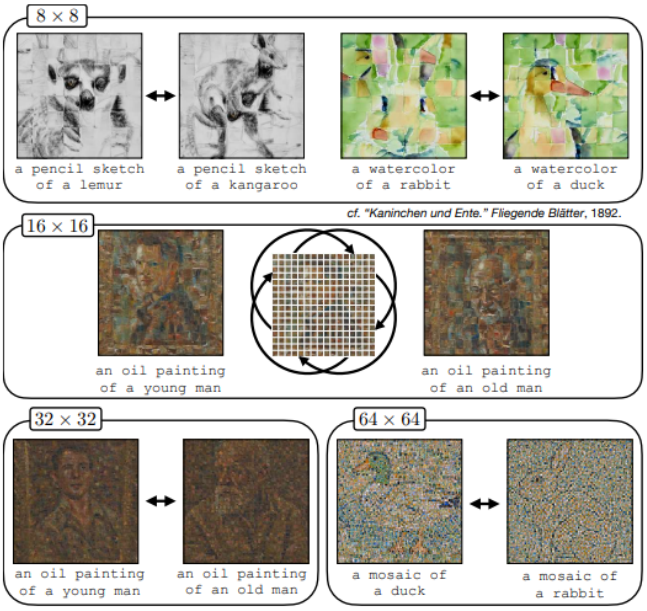

Figure 6. **Permutation Illusions.** We synthesize images whose appearance changes upon permutation of patches. Even in the difficult case of a 64 × 64 grid of patches, in which every pixel is effectively shuffled, we are able to generate meaningful images.

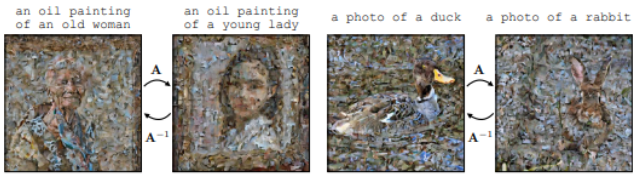

Figure 7. **Orthogonal Illusions.** We show that our method works, even when the view is a randomly sampled orthogonal transformation **A**. While these "illusions" are incomprehensible to human perception, they serve as a confirmation for our mathematical analysis.

Permutations（順列変換） ピクセルおよびパッチの順列変換は、直交変換の部分集合であり、本手法に適用可能です。図 6 では、ランダムにサンプリングした順列の下で、さまざまなサイズのパッチグリッドに対する結果を示します。64 × 64 の場合は非常に難しいものの、本手法は制約を満たす画像を生成できていますが、品質はやや低いです。

Arbitrary Orthogonal Transformations（任意の直交変換） セクション 3.3 で述べたように、本手法は任意の直交変換に適用可能です。これまでは、直交ビューの部分集合に基づく錯視を示してきましたが、これらは直感的な画像変換に対応しています。図 7 では、ビューとして任意の直交変換を使用した「錯視」を示します。この変換には、Stable Diffusion [36] を用い、SVD によって独立同分布なランダムガウス行列を射影して得られたランダムな直交行列 $A \in R^{16384\times16384}$ を使用しました。これらの次元は、Stable Diffusion の潜在空間のサイズに対応しています。この変換は画像にとって非常に困難で不自然ですが、本手法はそれでも合理的な画像を生成することができます。

Random Samples（ランダムサンプル） 選択したプロンプトに対するランダムサンプルを図 8 に示します。図 1 に示したものほど良くはありませんが、これらのランダムサンプルも非常に高品質であることが分かります。一部の失敗例として、モデルが 1 つのプロンプトを他よりも好む場合が見られます。さらなる議論とランダムサンプルについては付録 D を参照してください。

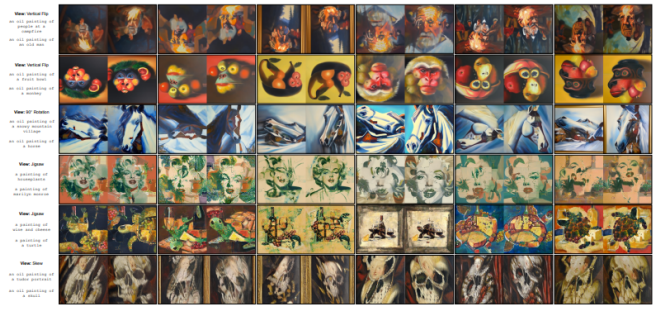

Figure 8. **Random Samples.** We show random samples, along with their corresponding view, for selected prompts. For more random samples please see Appendix D. **For best quality, view digitally and zoom-in.**

5.4　4.4 失敗例

本手法の興味深い失敗例を図 9 に示します。以下に、その具体的なケースを説明します。

独立した生成 (Independent Synthesis)　最初の失敗例は、モデルがプロンプトを別々に生成し、2 つのプロンプトを組み合わせて錯視を形成することに失敗するケースです。経験的に、このようなケースは驚くほど少ないことが分かっています。特に、これが簡単な近道となる解決策に思えるにもかかわらずです。我々は、この現象が拡散モデルがその内容を中央に配置することに偏っているためと仮定しています。このため、統合され、中央に配置された内容を持つ画像の方が、分離され、中央から外れた画像よりも多く生成されます。

ノイズの変化 (Noise Shift)　ノイズの統計を維持するビューを使用することは、本手法の成功にとって非常に重要です。例えば、「ドレス錯視」[46]（青と黒、または白と金のドレスとして見える錯視）を再現しようとした場合、ピクセル値を一定の係数でスケーリングする単純なホワイトバランス処理をビューとして使用しました。この変換は線形ではありますが、ガウスノイズの統計を維持しません。その結果、ノイズのスケーリングピークを信号として解釈し、積極的にノイズを除去することでスポット状のアーティファクトが現れます。

相関ノイズ (Correlated Noise)　本手法は変換として回転をサポートしていますが、図 1 に示す「3 ビュー」「4 ビュー」「内部回転 (Inner Rotation)」錯視で実証されているように、回転がノイズに相関を導入しないよう注意が必要です。例えば、双線形サンプリングでは 4 つの隣接するピクセルの線形結合としてノイズに大きな相関を導入します。そのため、一見無害な回転でも、変換が慎重に相関を排除していない場合、分散したサンプルが生成される可能性があります。図 9 の 45 度の双線形回転では、このような問題が発生していることが示されています。

6　5. 制限事項と結論

本研究では、魅力的かつ多様な光学的錯視を生成する手法を提案しました。本手法はシンプルで実装が容易であり、理論的な解析にも適しています。本手法が広範な変換セットに対して機能することを証明し、幅広い光学的錯視を生成できることを定性的に示しました。

しかしながら、本手法では未対応の錯視や変換も多く存在します。例えば、色の恒常性に関する錯視、射影変換（ホモグラフィー）、ストレッチ、さらにはより一般的な体積保存しない変形などが挙げられます。これらのビューの実装は今後の課題とします。

さらに、本手法は完全な錯視を一貫して生成できるわけではありません。これは良い錯視を生成することの困難さを反映している可能性もありますが、一貫性を向上させるためのさらなる研究の必要性を示している可能性もあります。

謝辞

早期の草稿に対するフィードバックを提供してくださった William Henning、Trenton Chang、Kimball Strong、Jeong-soo Park、Patrick Chao、Kurtland Chua、Mohamed El Banani に感謝します。また、Daniel は National Science Foundation Graduate Research Fellowship（Grant No. 1841052）の支援を受けています。

Figure 9. **Failures.** We highlight three interesting failure cases, which are discussed in Sec. 4.4.

参考文献

1. AUTOMATIC1111. Negative prompt. `https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative` 2022. Accessed: November 7, 2023. 3, 5

2. Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. `https://ryanndagreat.github.io/Diffusion-Illusions`, 2023. 2, 3, 5, 6

3. Kartik Chandra, Tzu-Mao Li, Joshua Tenenbaum, and Jonathan Ragan-Kelley. Designing perceptual puzzles by differentiating probabilistic programs. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

4. Ming-Te Chi, Chih-Yuan Yao, Eugene Zhang, and Tong-Yee Lee. Optical illusion shape texturing using repeated asymmetric patterns. *The Visual Computer*, 30:809–819, 2014.

5. Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM Trans. Graph.*, 29(4):51–1, 2010. 2

6. Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

7. Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019. 2, 4

8. Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647, 2020.

9. Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International Conference on Machine Learning*, pages 8489–8510. PMLR, 2023. 2, 4

10. Werner Ehm. A variational approach to geometric-optical illusions modeling. *Proceedings of Fechner Day*, 27(1):41–46, 2011. 2

11. Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018. 3

12. William T Freeman, Edward H Adelson, and David J Heeger. Motion without movement. *ACM Siggraph Computer Graphics*, 25(4):27–30, 1991. 2

13. Timur Garipov, Sebastiaan De Peuter, Ge Yang, Vikas Garg, Samuel Kaski, and Tommi Jaakkola. Compositional sculpting of iterative generative processes. *arXiv preprint arXiv:2309.16115*, 2023. 2, 4

14. Alexander Gomez-Villa, Adrian Martin, Javier Vazquez-Corral, and Marcelo Bertalmio. Convolutional neural networks can be deceived by visual illusions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12309–12317, 2019. 2, 3

15. Alex Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesus Malo. On the synthesis of visual illusions using deep generative models. *Journal of Vision*, 22(8):2–2, 2022. 2

16. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

17. Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. 2

18. Rui Guo, Jasmine Collins, Oscar de Lima, and Andrew Owens. Ganmouflage: 3d object nondetection with texture fields. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

19. Aaron Hertzmann. Visual indeterminacy in gan art. In *ACM SIGGRAPH 2020 Art Gallery*, pages 424–428. 2020. 2

20. Elad Hirsch and Ayellet Tal. Color visual illusions: A statistics-based computational model. *Advances in neural information processing systems*, 33:9447–9458, 2020. 2

21. Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 3

22. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2, 3

23. Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. *arXiv preprint arXiv:2309.16779*, 2023. 2, 3

24. Mikhail Konstantinov, Alex Shonenkov, Daria Bakshandaeva, and Ksenia Ivanova. If by deepfloyd lab at stabilityai, 2023. GitHub repository. 2, 5, 11

25. Monster Labs. Controlnet qr code monster v2 for sd-1.5, 2023. 3

26. Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, and Antonio Torralba. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34:23166–23178, 2021. 2, 4

27. Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 2, 4

28. Dominique Makowski, Zen J Lau, Tam Pham, W Paul Boyce, and SH Annabel Chen. A parametric framework to generate visual illusions using python. *Perception*, 50(11):950–965, 2021. 2

29. Jerry Ngo, Swami Sankaranarayanan, and Phillip Isola. Is clip fooled by optical illusions? 2023. 2, 3

30. Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021. 2

31. Aude Oliva, Antonio Torralba, and Philippe G. Schyns. Hybrid images. *ACM Trans. Graph.*, 25(3):527–532, 2006. 2

32. Andrew Owens, Connelly Barnes, Alex Flint, Hanumant Singh, and William Freeman. Camouflaging an object from many viewpoints. 2014. 2

33. Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3

34. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

35. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

36. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 7

37. Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

38. Troy Shinbrot, Miguel Vivar Lazo, and Theo Siu. Network simulations of optical illusions. *International Journal of Modern Physics C*, 28(02):1750018, 2017. 2

39. Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 2, 3

40. Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 2, 3

41. Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2, 3

42. Matthew Tancik. Illusion diffusion. `https://github.com/tancik/Illusion-Diffusion`, 2023. 2, 3, 5, 6, 12

43. Ugleh. Spiral town - different approach to qr monster. `https://www.reddit.com/r/StableDiffusion/comments/16ew9fz` 2023. 3

44. Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3

45. Xi Wang, Zoya Bylinskii, Aaron Hertzmann, and Robert Pepperell. Toward quantifying ambiguities in artistic images. *ACM Transactions on Applied Perception (TAP)*, 17(4):1–10, 2020. 2

46. Wikipedia contributors. The dress. `https://en.wikipedia.org/wiki/The_dress`. Accessed: November 9, 2023. 8

47. Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

付録 A: 実装の詳細

我々は、DeepFloyd IF [**?**] 拡散モデルの最初の 2 つのピクセルベースのステージを使用しました。具体的には、解像度が 64×64 の画像を生成する第 1 ステージと、画像を 256×256 にアップサンプリングする第 2 ステージを使用しています。我々の手法は両方のステージで適用され、各解像度に対してビュー変換を実装しています。DeepFloyd IF はさらに、ノイズ推定に加えて分散も予測します。複数の分散推定値を平均化することで統合します。ガイダンスの

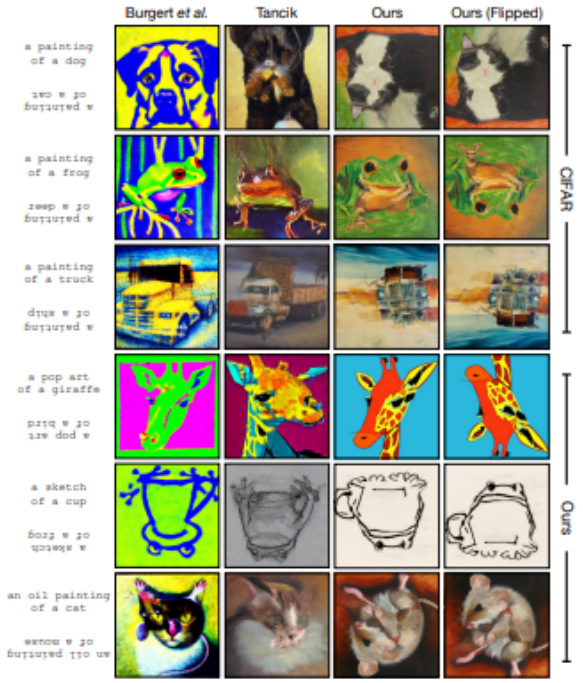

Figure 10. **Qualitative Comparisons.** We compare more illusions generated by baselines to our illusions. We show examples from both our prompt dataset and the CIFAR prompt dataset.

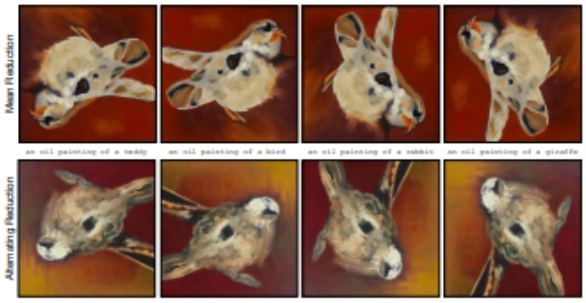

Figure 11. **Combining Noise Estimates.** We show that mean reduction does better than alternating with an example of a 4-view sample image.

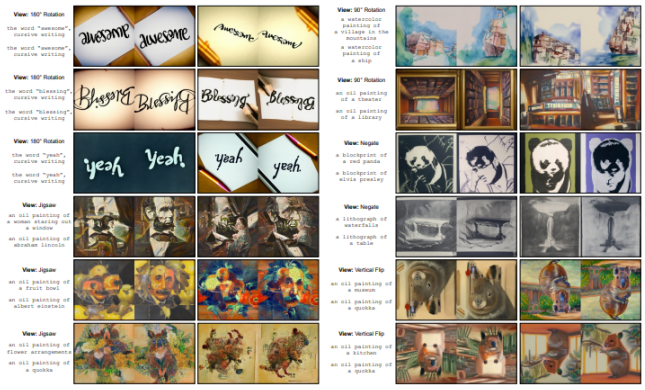

Figure 12. **Qualitative Samples.** We show more illusions with views such as rotations, flips, color inversion, and jigsaw puzzles.

強度は 7 から 10 の範囲で設定し、プロンプトに応じて 30 から 100 の推論ステップを使用します。両ステージで M サイズのモデルを使用しました。

DeepFloyd IF は分散も推定するため、これらの分散推定値にも逆変換を適用する必要があります。ピクセルの置換ベースのビューでは、単純に分散推定値に逆置換を適用します。一方、反転変換においては、予測された対数分散を反転することは意味をなしません。この場合、分散推定値を単純に反転しない方法が適切であることが分かりました。

DeepFloyd IF はさらに第 3 の超解像ステージを持ち、これは Stable Diffusion 4× アップスケーラーです。このモデルは 256×256 から 1024×1024 へのアップスケールを行います。このモデルは潜在空間モデルであるため、我々の手法を適用していません。ただし、アイデンティティビューに関連付けられたプロンプトでアップサンプリングを行うことで、各ビューの品質を損なうことなく使用可能であることが分かりました。Fig. 1 のすべての結果は、この方法でアップサンプリングされています。

付録 B: データセットの収集

我々のデータセットは、「a street art of...」や「an oil painting of...」といったスタイルのリストと、「an old man」や「a snowy mountain village」といった対象物のリストで構成されています。対象物とスタイルは手作業で選定され、GPT-3.5 を参考にしました。プロンプトペアは、スタイルプロンプトをランダムにサンプリングし、それを 2 つのランダムに選ばれた対象プロンプトに追加することで生成されます。

CIFAR データセットは、CIFAR-10 の 10 クラスを対象物として使用し、スタイルプロンプトとして「a painting of」を使用して構成しました。すべての 45 ペアの対象物を取り、それぞれにスタイルプロンプトを追加することで、45 個のプロンプトペアを作成しました。

付録 C: 追加結果

本セクションでは、追加の定性的結果を提供します。Fig. 10 では、我々の手法をデータセットおよび CIFAR プロンプトデータセットを使用したベースラインと比較しています。これは Fig. 5 の拡張です。また、90° と 180° の回転、アンビグラム、「ポリモーフィック」ジグソーパズル、色の反転、垂直反転を用いたさらなる錯視も生成しました（Fig. 12 参照）。さらに、同じ反転プロンプトと異なる非反転プロンプトを用いた複数の反転錯視を Fig. 13 に生成し、これらの反転版を Fig. 14 に示しています。

付録 D: ランダムサンプル

我々の手法を用いて生成されたさらなるランダムサンプルを提供します。回転、色反転、垂直反転に関しては Fig. 16 を参照してください。また、三視点、内部回転、「ポリモーフィック」ジグソーパズル、パッチおよびピクセル置換ビューについては Fig. 17 を参照してください。さらに、CIFAR データセットからのプロンプトを用いて生成されたランダムサンプルを Fig. 15 に示します。

CIFAR プロンプトペアの結果（Fig. 15 および Tab. 1）は、ランダムなプロンプトの代理として含まれています。ただし、完全にランダムなプロンプトを評価するために系統的にサンプリングするのは難しいことに注意が必要です。第一に、ランダムプロンプトをサンプリングする標準的な方法が存在しません。第二に、すべてのプロンプトペアが

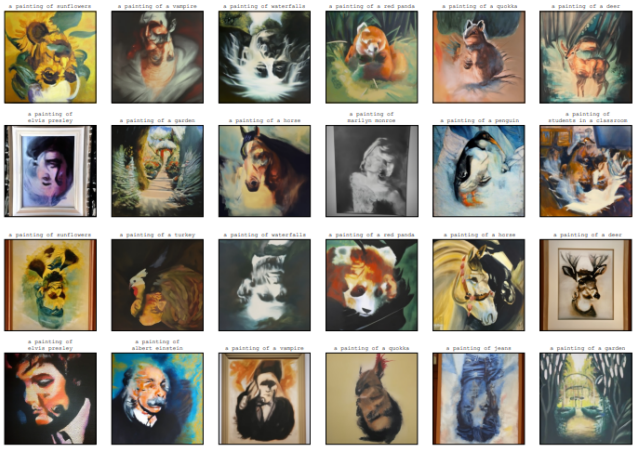

Figure 13. **Flip illusions.** For each row, the prompt of the flipped image is the same. We encourage the reader to guess what the flipped prompt is. For an answer and flipped illusions, please see Fig. 14.

良い錯視を作るとは限りません。例えば、スタイルが異なるプロンプトペアは錯視を生成するのに適していないことがあります。このため、Fig. 8、Fig. 16、Fig. 17 に示されるプロンプトは、ある程度キュレーションされたものです。

付録 E: プロンプト選択のアート

良い錯視を達成するためには、適切なプロンプトを選ぶことが重要であると分かりました。ここではいくつかの指針を示します。

まず、何が良い錯視を作るのかを推測するのは非常に難しいです。一見してうまくいきそうなプロンプトが一貫して失敗する場合もあれば、全く成功しそうにないプロンプトが驚くほど良く機能する場合もあります。我々は、抽象的なスタイル（例えば「a painting」や「a drawing」）が現実的なスタイル（例えば「a photo of」）よりもはるかに良く機能することを見出しました。これは、現実的なスタイルの制約が強すぎて、錯視がうまく機能しないためだと考えています。また、人間の顔は良い錯視を作るのに適していることも分かりました。これは、人間の視覚システムが顔のような刺激に対して特に敏感であるためだと推測されます。

付録 F: ジグソーパズルの実装

ジグソーパズルを生成するために、ピクセルの置換としてパズルピースの再配置を実装しました。まず、手作業で 3 種類のパズルピース（コーナーピース、エッジピース、センターピース）を描きました。これらのピースは、64 × 64、256 × 256、または 1024 × 1024 の画像を無重複でタイル化できるように設計されています。パズル内のすべてのピースは、これら 3 種類のピースのいずれかであり、それぞれ異なる向きを持ちます。次に、コーナー、エッジ、センターのピースそれぞれについてランダムな置換をサンプリングし、このピースの置換をピクセルの置換に変換しました。

付録 G: ノイズ推定値の統合

ノイズ推定値の平均を取る代わりに、タイムステップごとにノイズ推定値を交互または循環させる方法も試しました([42] 参照)。しかし、この方法では「スラッシング」が発生することがあります。これは、サンプルが異なるタイムステップで異なる方向に最適化され、品質の低下を招く現象です。さらに、2 つ以上のビューを持つ錯視では、各ビューのデノイズステップが減少し、品質の低下につながります。

例えば、4 つのプロンプト（「a teddy」、「a bird」、「a rabbit」、「a giraffe」）をそれぞれ画像の回転に対応させた場合、平均統合の方法は、Fig. 11 に示されるように、交互統合の方法よりも高品質な画像を生成します。

付録 H: ビューの線形性

Sec. 3.3 で述べたように、ビュー v が線形変換である場合、以下を満たします：

$$v(x_t) = v(w^t_{\text{signal}} x_0 + w^t_{\text{noise}} \epsilon) \tag{10}$$
$$= w^t_{\text{signal}} v(x_0) + w^t_{\text{noise}} v(\epsilon). \tag{11}$$

これは便利であり、ノイズ付き画像 x_t に v を適用することが、信号 x_0 とノイズ ϵ にそれぞれ独立に v を適用することと同等であることを意味します。また、結果は変換された信号と変換されたノイズの線形結合であり、タイムステップ t に対して拡散モデルが期待する重み付けとなります。

しかし、他の条件が成立する可能性もあります。例えば、以下を満たすことを仮定できます：

$$v(x_t) = v(w^t_{\text{signal}} x_0 + w^t_{\text{noise}} \epsilon) \tag{12}$$
$$= w^t_{\text{signal}} v_1(x_0) + w^t_{\text{noise}} v_2(\epsilon), \tag{13}$$

ここで、v が何らかの方法で信号とノイズに対して異なる方法で作用し、それを v_1 と v_2 を通じて結合し、適切な重み付けを行うという解釈が可能です。この点については今後の研究課題とします。

付録 I: 統計的一貫性

Sec. 3.3 で述べたように、$\epsilon \sim \mathcal{N}(0, I)$ および正方行列 A について、$A\epsilon \sim \mathcal{N}(0, I)$ が成り立つための必要十分条件は A が直交行列であることです。この証明を以下に示します。

ガウス分布の性質より、$A\epsilon$ もガウス分布であるため、平均と共分散を計算するだけで十分です。

平均は以下のようになります：

$$E[A\epsilon] = AE[\epsilon] = 0. \tag{14}$$

平均が 0 であるため、共分散は以下で与えられます：

$$\text{Cov}(A\epsilon) = E[(A\epsilon)(A\epsilon)^\top] \tag{15}$$
$$= AE[\epsilon\epsilon^\top]A^\top \tag{16}$$
$$= AA^\top. \tag{17}$$

したがって、$A\epsilon \sim \mathcal{N}(0, I)$ である場合、共分散 $\text{Cov}(A\epsilon) = AA^\top = I$ を満たさなければならず、これは A が直交行列であることを意味します。また、A が直交行列である場合、$AA^\top = I$ が成り立つため、$A\epsilon \sim \mathcal{N}(0, I)$ が成り立ちます。