IllusionVQA: 視覚と言語モデルのための挑戦的な光学的錯覚データセット

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee,
Wasi Uddin Ahmad, Yue Dong, Rifat Shahriyar
バングラデシュ工科大学, AWS AI Labs, カリフォルニア大学リバーサイド
sameen2080@gmail.com, salkhon050@gmail.com

Abstract

視覚と言語モデル（VLM）の登場により、研究者は自然言語を使用してニューラルネットワークの視覚的理解を調査することが可能になりました。物体の分類や検出を超えて、VLM は視覚的理解や常識的推論も可能です。この進展は、「画像自体が本質的に非合理的である場合、VLM はどのように反応するのか？」という問いを自然に生じさせました。この目的のために、私たちは IllusionVQA を提案します。このデータセットは、多様で挑戦的な光学的錯覚や解釈が困難なシーンを収録しており、VLM の能力を次の 2 つの異なる選択式 VQA タスクでテストすることを目的としています：理解タスクとソフトローカリゼーションタスクです。

最先端の VLM である GPT4V は、理解タスクで 62.99%の精度（4 ショット）を達成し、ローカリゼーションタスクでは 49.7%の精度（4 ショットおよび Chain-of-Thought）を示しました。一方で、ヒューマン評価では、理解タスクで 91.03%、ローカリゼーションタスクで 100%の精度を達成しました。

また、In-Context Learning（ICL）と Chain-of-Thought 推論がローカリゼーションタスクにおいて Gemini-Pro の性能を大幅に劣化させることを発見しました。さらに、VLM の ICL 能力には潜在的な弱点があることが分かりました。それは、正解が数ショットの例としてコンテキストウィンドウ内に存在している場合でも、光学的錯覚を特定できないという問題です。

1 1 Introduction

光学的錯覚は、現実とは異なるとされる視覚的な知覚によって特徴付けられる。錯覚にはさまざまな種類があり、その分類は難しい。なぜなら、その原因がしばしば明確でないためである（Bach & Poloschek, 2006; Gregory, 1997a）。図 2 では、互いに異なるがすべて光学的錯覚の範疇に入る例を示している。図 2a は不可能物体を示し、「物体の一部（この場合は象の脚）が背景となり、その逆も然りである。」（Shepard, 1990）。図 2b は 3 × 3 のグリッドを描写しており、一部のセルが淡黄色に見えるが、よく観察するとすべてのセルは完全に白色であることが分かる。図 2c は、白い立方体の上にカップが休んでいるかのように配置されている、現実的な 3D 描写を示している。これら 3 つの例はすべて光学的錯覚とされるが、共通点はほとんどない。

光学的錯覚の分類における課題にもかかわらず、Gregory（1997b）は外見に基づいて以下の 4 つの主なクラスを提案した：曖昧さ、歪み、逆説、および虚構。このフレームワークに基づくと、図 2a は逆説に該当する。なぜなら、描写された物体は実際には存在し得ないからである。図 2b は歪みであり、いずれの領域も淡黄色ではないにもかかわらず、そのように見えるからである。図 2c はトロンプ・ルイユ（だまし絵）の一例であり、その分類が光学的錯覚として議論されているものの（Wade & Hughes, 1999）、上述の虚構カテゴリに該当する可能性がある。

光学的錯覚に対する私たちの知覚の原因は、認知心理学の豊富な研究分野である（Helmholtz, 1948; Gregory, 1968; Gentilucci et al., 1996）。人間の視覚認知と人工ニューラルネットワークの類似性（Cichy et al., 2016; Eickenberg et al., 2017）に動機づけられ、研究者たちは人工ニューラルネットワークが光学的錯覚をどのように知覚するかを探求してきた（Gomez-Villa et al., 2019; 2020; Sun & Dekel, 2021; Lonnqvist et al., 2021; Hirsch & Tal, 2020）。しかし、これらの研究の範囲と一般化可能性は、個々の光学的錯覚に対するモデルの活性化や勾配をケースバイケースで分析する必要性によって制限されていた。この問題を解決するため、Zhang et al. (2023) は、光学的錯覚を小型の VLM（最大 130 億パラメータ）を通じて初めて探求し、自然言語を用いて光学的錯覚に関する内部的な信念を調査した。

これまでの研究とは異なり、本研究ではインターネットから収集した挑戦的な光学的錯覚を、認知心理学の研究に基づいた 12 の異なるカテゴリに分類した。さらに、詳細な質問と回答のペアを作成し、VLM（Vision Language Models）の能力を調査するために設計されたいくつかの誤答選択肢を含めた。我々は、この問題を標準的な視覚的質問応答タスク（VQA）として定式化し、最先端の VLM（例えば GPT4V（Achiam

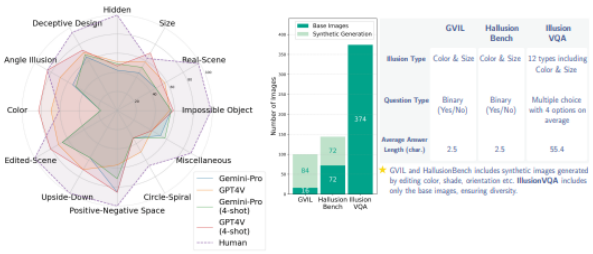


Figure 1: Left: Comparison of human and VLM performance in IllusionVQA-Comprehension. Right: Comparison of IllusionVQA with prior illusion datasets - GVIL (Zhang et al., 2023) and HallusionBench (Liu et al., 2023a).

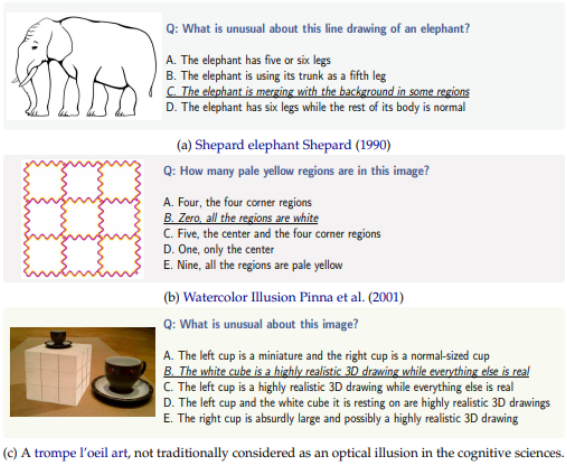


Figure 2: Examples of optical illusions in IllusionVQA-Comprehension

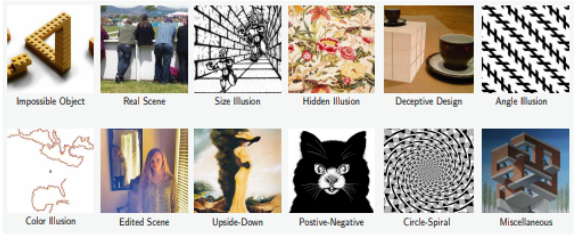


Figure 3: Categories in IllusionVQA-Comprehension. Refer to Appendix D for details.

et al., 2023）や Gemini-Pro（Team et al., 2023））および小型のオープンソースモデルを評価した。詳細な分類と広範な人間による評価を通じて、光学的錯覚に関する人間と VLM の認知をさまざまな軸で精緻に比較できる。

本研究の貢献は以下の通りである。1. IllusionVQA という新しいデータセットを導入し、VLM が挑戦的な光学的錯覚を見つけ、理解する能力を厳密にテストした。2. オープンソースおよびクローズドソースの幅広い VLM を包括的にテストした。実験の結果、GPT4V が錯覚の理解および位置特定において最も優れたモデルであることが判明したが、それでも人間のパフォーマンスには大きく及ばない。3. 実験を通じて、最先端の VLM が通常の物体を正確に特定できる一方で、光学的錯覚には苦戦することが明らかになった。さらに、これらのモデルは In-Context Learning や Chain-of-Thought 推論による評価で一貫性を欠くことが観察された。

2 Related Work

研究者たちは、通常の画像で訓練された畳み込みニューラルネットワークが特定の光学的錯覚に対して人間と同様に感受性を示すことを明らかにしている（Gomez-Villa et al., 2019; 2020; Afifi & Brown, 2019; Benjamin et al., 2019; Sun & Dekel, 2021）。しかし、これまでの多くの実験は光学的錯覚の特定のカテゴリに焦点を当てており、その手法はすべての種類に一般化できるものではなかった。

Zhang et al. (2023) は、自然言語を通じて VLM（Vision Language Models）が光学的錯覚をどのように認識するかを調査した初めての研究である。この研究では 16 枚の光学的錯覚の基底画像を収集し、手動編集を通じて 100 種類のバリアントを作成した。そして、モデルが光学的錯覚に「欺かれた」かどうかを

テストするため、Yes/No 形式の質問を設定した。複数のオープンソース VLM（最大 13B パラメータ）をテストした結果、大規模な VLM ほど光学的錯覚に対して感受性が高いことが判明した。

Liu et al. (2023a) は、光学的錯覚および幻覚に対する VLM の感受性をテストするためのベンチマークである *HallusionBench* を提案した。このベンチマークには、主に Zhang et al. (2023) から派生した 72 枚の光学的錯覚の基底画像が含まれている。Bitton-Guetta et al. (2023) は、「密閉されたボトルの中に立てられたろうそく」といった奇妙な画像に対する VLM の理解力をテストした。この研究では、テキストから画像を生成するモデル（Ramesh et al., 2021; Rombach et al., 2022）を使用して 500 枚の画像を生成し、画像が「奇妙」である理由を記述するために人間のアノテーターを雇用した。この結果、GPT3 は正解の画像説明が提供された場合に 68%の精度を達成したのに対し、人間は 95%の精度を達成した。

これまでの研究では、合成的な光学的錯覚生成のためのさまざまなツールが提案されてきた。Hirsch & Tal (2020) および Fan & Zeng (2023) は、それぞれ色に関連する錯覚および格子に関連する錯覚を自動生成する手法を開発した。Gomez-Villa et al. (2022) は GAN を使用して錯覚を生成する方法を探求している。しかし、合成的な錯覚生成の主な制約は、作成できる錯覚の種類が限られている点である。特に、文献で発見されている幅広い光学的錯覚と比較するとそのバリエーションは乏しい。そのため、既存の合成錯覚生成アルゴリズムにのみ依存して VLM を評価することは、その評価の範囲を著しく制限してしまう。

これに対し、本研究ではデータの多様性を確保するためにインターネットから光学的錯覚の画像を収集し、複数の軸に沿って VLM をテストした。我々は手動で複数選択式の質問応答データセットを作成し、それぞれの質問が曖昧さのない唯一の正解を持つように設計したが、他の選択肢はもっともらしく見えるものの誤りである。理解力を超えて、VLM が画像内の幾何学的に不可能な物体をどれだけ正確に特定できるかもテストした。我々の実験は、光学的錯覚の文脈で VLM が言語の事前知識（Goyal et al., 2017）および特徴エンタングルメント（Tang et al., 2023）にどれほど依存しているかを調査している。

3 The IllusionVQA Dataset

IllusionVQA は、2 つのサブタスクを含む視覚的質問応答（VQA）データセットです。最初のタスクでは、光学的錯覚 12 カテゴリの 435 例において理解力をテストします。各例は、光学的錯覚の画像、質問、および 3〜6 の選択肢から構成され、そのうち 1 つが正しい答えとなっています。このタスクを **IllusionVQA-Comprehension** と呼びます。2 番目のタスクでは、VLM が幾何学的に不可能な物体を通常の物体と区別できる能力をテストします。このタスクは、最初のタスクと同様の形式で 1000 の例を含んでいます。このタスクを **IllusionVQA-Soft-Localization** と呼びます。本節では、画像収集とフィルタリングについて述べた後、Comprehension タスクの QA 生成と分類、Soft-Localization タスクの手続き的 QA 生成について説明します。

3.1 Image Collection and Filtering

我々は、複数のオンラインリポジトリから光学的錯覚の画像を 3500 枚以上収集しました。それらの画像を手動で確認し、光学的錯覚であることを確認しました。VLM はウェブから収集したデータで訓練されているため、ウェブ収集画像データセットにおけるデータ汚染は大きな懸念事項です。Liu et al. (2023a) によれば、GPT4V はすべての錯覚ケースを認識し、それらの名前を知っています。我々は、現実世界の光学的錯覚の多様性を捉えるため、合成生成ではなくウェブデータを選びました。

データ汚染のリスクを軽減するために、GPT4V を使用してデータセットをフィルタリングしました。具体的には、GPT4V に画像を説明させ、その応答を検査しました。GPT4V が光学的錯覚を正確に検出し説明できた場合、その錯覚を IllusionVQA に含めませんでした。一部のケースでは、GPT4V が光学的錯覚の存在を検出できたものの、それを正確に説明できなかった（例えば、不可能な立方体（Penrose & Penrose, 1958）を Necker 立方体（Rosenholtz, 2011）と誤解した）ことがありました。これらのケースは IllusionVQA に含めました。

最終的に、高品質な光学的錯覚画像 374 枚が残りました。これらの画像はすべて、最新の GPT4V および Gemini-Pro API の内部フィルタチェックを通過しています。各画像の出典は、データセットリポジトリに含まれています。フィルタリングから除外された画像のいくつかの例を**付録 I** で示しています。

3.2 QA Generation for IllusionVQA-Comprehension

我々は、IllusionVQA-Comprehension タスクを、(Gurari et al., 2018; Goyal et al., 2017) と類似した標準的な選択肢形式の VQA 設定として提示します。一部の例では、錯覚が存在することを認識するだけで十分です（図 2a）。他のケースでは、VLM に対して画像に関する具体的な質問を行い、錯覚に関する理解を探ります（図 2c）。

光学的錯覚の単なる検出を超えて、VLM の理解を探る適切な質問を作成します。IllusionVQA 内の各質問には 1 つの正しい答えと複数の誤った選択肢があります。質問と正答を作成し、明確に間違った代替選

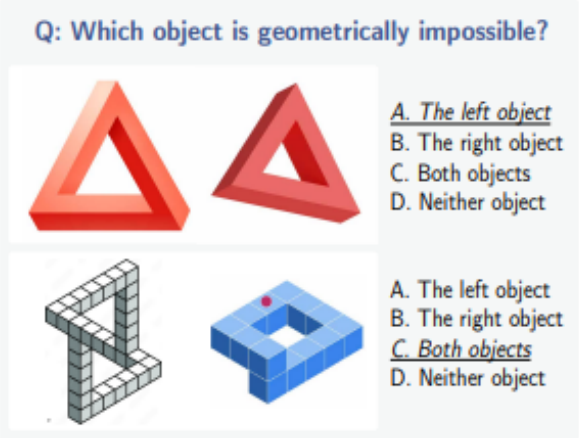


Figure 4: Examples demonstrating the task of Soft-Localization in IllusionVQA.

択肢を含めます。これにより、曖昧さや不明確さを排除します。誤答選択肢の生成には以下の方法とヒューリスティックを使用しました：

1. 光学的錯覚の最も可能性の高い誤解を可能な限り代替選択肢として含めました（例：図 2b の「5、中央と四隅の領域」という選択肢）。
2. 画像、質問、異なるモデルで生成された誤った記述を代替選択肢として VLM に提示しました。
3. VLM が視覚入力を無視しがちであることを考慮し（Goyal et al., 2017; Jabri et al., 2016）、画像なしで質問のみを提示した場合に生成される誤った記述も代替選択肢として含めました。
4. 各誤答選択肢を手動で検査・編集し、質問と画像を考慮して難易度の高いものにしました。

適切な場合には 1 つの画像に対して複数の質問を作成し、それらの質問が単なる言い換えにならないよう注意しました。また、Perspectives API[1]を使用して、有害なコンテンツをフィルタリングしました。

結果として、慎重に選定された 374 枚の光学的錯覚画像から 439 の質問と回答のペアを作成しました。厳格なフィルタリングと質問・回答生成技術を通じて、IllusionVQA データセット内のすべてのインスタンスが非常に挑戦的でありながら明確であることを保証しました。我々のデータセットには、GVIL (Zhang et al., 2023) や HallusionBench (Liu et al., 2023a) に存在する画像と似た 30 枚の画像が含まれていますが、尋ねられる質問は異なります。さらに、我々のデータセットでは、GVIL や HallusionBench のブール QA ではなく、選択肢形式の QA を通じて VLM を評価します。

3.3 IllusionVQA-Comprehension Classification

Gregory (1997b) および Bach & Poloschek (2006) によって提案された分類方法を本データセットに適応し、図 3 に示す 12 クラスに分類しました。重み、勾配、および活性化へのアクセスなしに VLM をブラックボックス設定で研究するという目標により適合する外観ベースの分類を採用しました。光学的錯覚の分類に関して明確な合意はないため、クラス間にある程度の重複が避けられませんでした。特に、Deceptive Design カテゴリは Real Scene カテゴリのサブセットと見なされる可能性があり、Edited Scene カテゴリも Impossible Objects カテゴリに含まれる場合があります。しかし、画像は十分に視覚的に異なっているため、別々の分類が必要であると判断しました。

さらに、Deceptive Design、Edited Scene、および Upside Down カテゴリは、認知科学において一般的に光学的錯覚とは見なされていません。しかし、これらの画像は日常的には光学的錯覚と見なされることが多いため、これらの画像を含めました。光学的錯覚とその分類に関する関連文献については付録 A および C で議論します。

3.4 IllusionVQA-Soft-Localization のための手続き的 QA 生成

我々は「不可能物体」（図 3 参照）の画像を手続き的に組み合わせることで、VLM がシーン内の錯覚を特定できるかどうかをテストします。VLM は境界ボックスの生成において優れた能力を示してきました（Bai

[1]Perspectives API は、質問、選択肢、および回答内の有害な内容をフィルタリングするために使用しました。

et al., 2023; Wang et al., 2023）が、GPT4 や Gemini-Pro はこのタスク向けに設計されていません。そのため、幾何学的に不可能なオブジェクトを横並びに配置した 2 つの画像を使用し、VLM に「不可能物体の位置」を特定させるという、より簡単なローカライゼーションタスクを採用しました。このタスクを Soft-Localization と呼びます。ここで、モデルは「左」または「右」といった回答を行うだけで済みます（Precise Localization のように正確な境界ボックス座標を出力する必要はありません）。

Soft-Localization の能力を評価するため、幾何学的に不可能なオブジェクト 40 点と、通常の幾何学オブジェクト 20 点の図を収集しました。また、外観が似た通常および不可能なオブジェクトをグループ化しました。これらを用いて、4 つの選択肢（A. 左側のオブジェクト、B. 右側のオブジェクト、C. 両方のオブジェクト、D. どちらでもない）に対する 250 枚の画像を生成し、合計で 1000 枚の画像を作成しました。この過程で、見た目が似ているペアを優先しました。その結果、IllusionVQA-Soft-Localization はラベルがバランスされた 1000 サンプルで構成されています。

幾何学的に不可能なオブジェクトの Soft-Localization は、輪郭を正確に検出することや、高度な幾何学的・空間的推論を必要とするため、VLM にとって難しい課題となります。制御実験を通じて、通常のオブジェクトを用いた Soft-Localization タスクでの VLM の能力を検証し、その結果を付録 F で示します。

4 実験設定

オープンソースの VLM は複数の画像入力をサポートしていないため、オープンソース VLM については 0-shot で評価を行い、クローズドソース VLM については 0-shot および 4-shot の設定で評価を行いました。IllusionVQA-Comprehension では、4 つの最も一般的な錯覚カテゴリからそれぞれ 1 つのサンプルを選択し、few-shot の例として使用しました。IllusionVQA-Comprehension のテストセットには 435 のインスタンス（370 枚の画像）が含まれています。光学的錯覚を自然言語で記述することの難しさから、理解タスクにおいて Chain-of-Thought（CoT）評価は試みませんでした。IllusionVQA-Soft-Localization では、4 つの選択肢のそれぞれについて 1 つのインスタンスを few-shot の例として選択しました。4-shot+CoT 評価の場合には、各例に対して標準化された推論テンプレート（付録 E 参照）を提供しました。デフォルトの API 引数と生成パラメータを使用しました。

すべての画像を 512 ピクセルにリサイズし、アスペクト比を維持しました。Soft-Localization タスクでは、オブジェクトの幾何学とは関係がないため、グレースケール変換を適用しました。執筆時点（2024 年 3 月）で利用可能な最新バージョンの GPT4V および Gemini-Pro を使用しました。また、GPT4o、Claude-3.5-Sonnet、その他の VLM の更新された結果（2024 年 7 月）を付録 J に含めました。オープンソース VLM としては、InstructBLIP（Dai et al., 2024）、LLaVA-1.5（Liu et al., 2023b）、CogVLM（Wang et al., 2023）の 3 つをテストしました。

人間による評価 3 人の人間の評価者を雇い、理解タスクのすべてのサンプルと、Localization タスクからランダムに選択された 200 のサンプルを提供しました。評価は WebUI を通じて実施され、各質問に費やされた時間をバックグラウンドで追跡しました。詳細は付録 G に記載しています。

5 結果

5.1 IllusionVQA-Comprehension

表 1 は、錯覚の理解における VLM の性能がすべてのカテゴリで人間の性能に大きく劣ることを示しています。ほとんどのカテゴリで大規模なモデルが小規模な VLM を上回り、GPT4V は 3 つのカテゴリを除いて他の VLM を上回っています。

人間の性能 IllusionVQA の選択式設定により、錯覚の記述の難易度が大幅に軽減されます。その結果、3 人の評価者の正答率は 85%を超え、多数決とランダムなタイブレイクを用いることで 91.03%に達しました。最も性能の高い VLM である GPT4V（4-shot）は、人間の評価者を「サイズ」と「色」の 2 つのカテゴリでのみ上回りました。評価者間の一致度と回答時間については付録 G で詳述しています。

GPT4V の優位性 GPT4V は、0-shot および 4-shot の両設定で他の VLM を大きく引き離しており、それでも人間の性能には遠く及びません。我々はさらに GPT4V の出力に対するケース分析を行いました。GPT4V は IllusionVQA の 12 カテゴリのうち 10 カテゴリでリードし、残りの 2 カテゴリでも競争力のある性能を示しました。「Real-Scene」、「Size」、「Color」、「Deceptive Design」などのカテゴリは、深度、色、そして挑戦的なシーンに対する VLM の認識をテストします。これらのスキルは、自律型ロボットとして VLM を現実世界で使用する際に特に関連性があります（Zitkovich et al., 2023; Wake et al., 2023）。GPT4V はこれら 4 カテゴリのすべてで大幅なリードを維持し、挑戦的な現実のシーンに対しても耐性を示し、Gemini-Pro に対して 10.93%のリードを保っています。

小規模 VLM 小規模な 3 つの VLM（パラメータ数 $\sim 14 - 17B$）の中では、LLaVA-1.5 と CogVLM がほとんどのカテゴリで類似した性能を示しました。一方、InstructBLIP は異なるカテゴリで最も大きな

Class	#	0-shot					4-shot		Human
		I-BLIP	LLaVA	Cog	Gemini	GPT4V	Gemini	GPT4V	
Impossible Object	134	34.22	43.28	44.03	**56.72**	55.22	56.72	58.96	98.51
Real-Scene	64	26.56	42.19	34.38	46.88	**57.81**	46.88	54.69	98.44
Size	46	26.09	19.57	13.04	45.65	**58.70**	52.17	69.57	63.04
Hidden	45	44.44	42.22	42.22	42.22	**51.11**	48.89	46.67	100.0
Deceptive Design	37	37.84	43.24	45.95	64.86	**70.27**	67.56	72.97	94.59
Angle Illusion	26	30.77	38.46	30.77	53.85	**69.23**	50.00	84.62	84.62
Color	23	30.43	26.09	30.43	17.39	**69.57**	17.39	82.61	60.87
Edited-Scene	21	42.86	61.90	42.86	66.67	**71.43**	66.67	80.95	100.0
Upside-Down	7	42.86	**71.43**	**71.43**	57.14	**71.43**	57.14	71.43	100.0
Pos.-Neg. Space	7	57.41	42.86	71.43	**85.71**	57.14	71.43	85.71	100.0
Circle-Spiral	6	33.33	0.00	16.67	33.33	**50.00**	33.33	33.33	66.67
Miscellaneous	19	36.84	42.11	42.11	**52.63**	42.11	57.89	42.11	89.47
Total	435	34.25	40.00	38.16	51.26	**58.85**	52.87	62.99	91.03

Table 1: Performance of VLMs on IllusionVQA-Comprehension. Categories where accuracy has improved using 4-shot prompting are underlined. 'Human' performance refers to the aggregated performance based on the majority vote of three human evaluators.

VLM	Orientation	Accuracy
LLaVA-1.5	Normal	80
Gemini-Pro	Normal	79.5
LLaVA-1.5	Vertically Flipped	73.5
Gemini-Pro	Vertically Flipped	64

Table 2: Performance of VLMs on a subset of 200 instances from VQA-v2.0 (Goyal et al., 2017). The model outputs were manually evaluated.

変動を示し、「サイズ」と「円-スパイラル」錯覚の理解において 5%以上の優位性を示す一方で、他のカテゴリでは大幅に低い性能を示しました。

5.2 In-Context Learning (ICL)

我々は GPT4V および Gemini-Pro を用いて 4-shot 学習をテストし、IllusionVQA-Comprehension における全体的な精度がわずかに向上することを観察しました。しかしながら、この向上はカテゴリ間で一貫していませんでした。例えば、4-shot GPT4V の精度は、「Real-Scene」、「Hidden Illusion」、および「Circle-Spiral Illusions」の 3 つのカテゴリで 0-shot の精度を下回りました。Gemini-Pro も同様の精度低下を示しましたが、低下したカテゴリは異なります。ICL は普遍的な戦略ではない可能性があり、few-shot の例が追加の言語的バイアスを導入し、VLM を誤った答えへと誘導する場合があることが示唆されます (Goyal et al., 2017; Jabri et al., 2016)。

5.3 通常の上下反転画像に対する VLM の性能

UpsideDown カテゴリは、縦方向の向きによって 2 つ以上の絡み合った要素を描写する画像を対象とします。表 1 に示されるように、小規模なオープンソース VLM はこのカテゴリで競争力のある性能を示しましたが、他の多くのカテゴリではクローズドソース VLM に大きく遅れを取っています。この矛盾を検証するために、我々は VQA-v2.0（Goyal et al., 2017）のサブセットを作成し、LLaVA-1.5 および Gemini-Pro を未編集および上下反転された画像でテストしました。

表 2 によれば、Gemini-Pro の精度は反転画像で 15.5%低下し、一方で LLaVA-1.5 の精度は 6.5%のみの低下に留まりました。この結果は表 1 での発見を裏付けるものです。

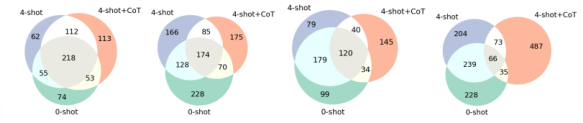
5.4 IllusionVQA-Soft-Localization

表 3 は、光学的錯視の位置特定が VLM にとって重大な課題である一方で、人間にとっては非常に容易であることを示しています。特に、小規模なオープンソース VLM はほぼランダムな性能（25%）を示しました。GPT4V および Gemini-Pro はやや良好な結果を示しましたが、人間の性能には及びませんでした。ランダムに選ばれた 200 のサンプルにおいて、3 人の人間評価者は全ての問題で正解を達成しました。ICL および CoT を使用しても、IllusionVQA-Soft-Localization データセットにおける GPT4V の性能は依然として低く、49.7%の精度しか達成しませんでした。[2]

[2]詳細は Appendix F を参照してください。

VLM	Prompt Type	Accuracy
InstructBLIP	0-shot	24.3
LLaVA-1.5	0-shot	24.8
CogVLM	0-shot	28
GPT4V	0-shot	40
	4-shot	46
	4-shot + CoT	49.7
Gemini Pro	0-shot	43.5
	4-shot	41.8
	4-shot + CoT	33.9
Human		100

Table 3: Performance of VLMs on IllusionVQA-Soft-Localization. Human performance denotes the majority-vote performance of three human evaluators on 200 random instances.



(a) GPT4 - correctly ans. (b) GPT4 - incorrectly ans. (c) Gemini-Pro - correctly ans. (d) Gemini-Pro - incorrectly ans.

Figure 5: Venn Diagrams showing the agreement between prompting techniques. There are instances where ICL and CoT cause the VLMs to answer incorrectly.

5.4.1 大規模 VLM は通常の物体を特定できるが錯視は特定できない

我々はソフトローカライゼーションのための新しいデータセットを作成しました。このデータセットは、異なるブランドのスポーツカーを使用し、「どの車が [BRAND] か？」という質問を同じ 4 つの選択肢（図 4 参照）で提示します。このタスクは、異なるブランドのスポーツカーが同じ基本形状を共有し、細部が異なるため、VLM にとって中程度に難易度の高い課題です。小規模な VLM はこのタスクで苦戦しましたが、GPT4V および Gemini-Pro は 100%の精度を達成しました。

GPT4V および Gemini が通常の物体のソフトローカライゼーションで優れた性能を示す一方で、IllusionVQA-Soft Localization での予想外の低性能は驚くべきことです。これは、幾何学的錯視を通常の物体と区別するためには、特徴抽出ではなく空間的推論が必要であるためと考えられます（空間的推論は VLM の弱点として知られています（OpenAI, 2023））。データセットとモデル性能の詳細については Appendix F を参照してください。

5.5 インコンテキストラーニング (ICL) とチェイン・オブ・ソート (CoT)

インコンテキストラーニング (ICL) (Brown et al., 2020) およびチェイン・オブ・ソート (CoT) (Wei et al., 2022) は、LLM の性能を向上させるために広く使用されている技術です。研究によれば、これらの技術はビジョンランゲージタスクにも効果的であることが示されています (Yang et al., 2023) が、十分に研究されているとは言えません。

表 3 は、Gemini-Pro が ICL や CoT を使用しない場合、GPT4V よりも高い性能を示し（+3.5%）、一方で GPT4V は 4-shot+CoT 評価時に全体で最高の精度を達成したことを示しています（+9.7%の向上）。対照的に、Gemini-Pro の性能は大幅に低下しました（-9.6%）。この原因として、GPT4V がより大規模なモデルであり、より強力な ICL 能力を発揮することが挙げられます (Wei et al., 2023)。

図 5 に示されるように、0-shot 評価で正解を得た一方で、4-shot+CoT 推論では失敗した事例が多数存在します。このことは、ICL および CoT が IllusionVQA-Soft-Localization に普遍的に適用可能ではないことを示唆しています。

5.6 インコンテキスト例における失敗

VLM の弱点を調査するために、追加の実験を行いました。この実験では、GPT4V および Gemini-Pro に対して 4 つの例（4-shot）を提示し、同じ例を質問として出題しました。この場合、VLM のコンテキストウィンドウ内に正確な画像、質問、および正解が含まれています。しかし驚くべきことに、GPT4V および Gemini-Pro のどちらも全ての質問に正しく回答することができませんでした。

これは、VLM におけるインコンテキストラーニング (ICL) の不可解な制限を浮き彫りにする結果です。具体的には、モデルがコンテキスト内に正しい答えが存在しているにもかかわらず失敗することが示されました。この現象は、言語の事前知識 (language priors) への過度の依存 (Goyal et al., 2017; Jabri et al., 2016) を指摘しており、視覚情報がほとんど無視されていることを示唆しています。この現象は、試行した全ての組み合わせで一貫して観察されました。このような組み合わせの一例についてのプロンプトと結果を付録 H に示しています。

6 6. Discussion

6.1 6.1 ロボティクスにおける錯覚耐性を備えた VLMs

Vision Language Models (VLMs) は、特にロボティクスの分野で現実世界のアプリケーションにおいてますます利用されています (Zitkovich et al., 2023; Wake et al., 2023)。これらのモデルは、ロボットが環境を認識し、対話する能力を高める上で重要な役割を果たしています。エンボディード AI は、誤解を招く設計、錯覚、隠されたオブジェクトを含む広範な現実世界の状況において、一貫した性能を示す必要があります。

IllusionVQA データセットは、エンボディード VLMs の堅牢性と適応性を評価する完全なストレステストとして機能する可能性を秘めています。この挑戦的なデータセットにおける VLMs の性能を評価することで、研究者は弱点を特定し、モデルが知覚的に曖昧なシナリオをナビゲートする能力を向上させることができます。

6.2 6.2 速い思考と遅い思考 - VLM と人間の反応時間の比較

Kahneman (2011) は、2 つの思考モード、「システム 1」（迅速で直感的）と「システム 2」（遅く、熟考的かつ論理的）を提唱しました。現在の最先端 LLMs は、その自己回帰型アーキテクチャにより、主に「システム 1」の思考を行うことが可能です (Bubeck et al., 2023)。LLMs で「システム 2」の推論を近似することは、Chain-of-Thought (Wei et al., 2022)、Tree-of-Thought (Yao et al., 2024) などの方法に関する重要な研究の動機となっています。

錯覚を理解し、それを特定するには、熟考的かつ論理的な「システム 2」の思考プロセスが必要です。人間の評価者は、IllusionVQA-Comprehension の各質問に平均 14.99 秒、IllusionVQA-Soft-Localization には平均 5.68 秒を費やしました。一方、GPT4V および Gemini-Pro の API は正確な推論時間を公開していないため、API 応答時間を報告します。0-shot (4-shot) 設定では、GPT4V は平均 1.81 秒 (3.27 秒)、Gemini-Pro は平均 3.82 秒 (4.59 秒) を要しました。API 応答時間はモデルの推論時間を大幅に過大評価するものの、それでも人間が各質問を熟考するのに費やす時間よりも著しく速いことがわかります。

7 Conclusion

現在の最先端 Vision Language Models (VLMs) が光学的錯覚の理解に苦戦し、幾何学的に不可能なオブジェクトの特定において大きく失敗することを示しました。我々の研究では、GPT4V がほとんどの錯覚タイプにおいて大きなリードを維持している一方で、オープンソース VLM とクローズドソース VLM の間に顕著な性能差が存在することが分かりました。また、GPT4V および Gemini-Pro が理解タスクにおいて数ショット学習が有効であることを検証しましたが、数ショット学習および Chain-of-Thought (CoT) 推論が局所化タスクにおいて普遍的に効果的ではないことも示しました。さらに、GPT4V および Gemini-Pro が正解が文脈内に提供されている場合でも、錯覚の局所化に関する質問に正しく答えられないことを発見しました。我々はデータセットと評価コードを公開し、さらなる研究を促進することを目指します。

8 Limitations

本研究は貴重な初期的な知見を提供しますが、いくつかの制限があり、それが将来の研究の課題を提示しています。

1. データセットの規模

IllusionVQA-Comprehension データセットの例の数は、インターネットから取得された 3500 枚以上の候補画像に厳格なフィルタリングプロセスを適用した結果、比較的少ないものとなっています。広範な努力にもかかわらず、我々の基準を満たす追加の光学的錯覚を見つけることができませんでした。将来的にデータセットを拡張するための潜在的な方法として、合成された光学的錯覚の利用が挙げられます。ただし、画

像生成モデルの現状では、新規で高品質な光学的錯覚を作成する能力が限られているため、現時点ではこのアプローチには困難が伴います。

2. 評価の範囲

OpenAI API のリクエスト制限（500 リクエスト/日）のため、GPT4V を複数回の評価実行でテストすることができませんでした。より高度な評価手法 (Lei et al., 2024) や、十分なデータを用いたオープンソース VLM のタスク特化型ファインチューニングによって、IllusionVQA-Soft-Localization の性能向上が期待できる可能性があります。

3. オープンエンド型質問応答

我々は、基準となる多肢選択型 VQA 設定の性能が十分でなかったため、主にオープンエンド型 QA の調査を控えました。オープンエンド型生成の複雑性は、評価者として追加の人間が必要となることでさらに増します。

References

1. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. *GPT-4 Technical Report.* arXiv preprint arXiv:2303.08774, 2023.

2. Edward H. Adelson. *Checker Shadow Illusion*, 1995. URL: `https://persci.mit.edu/gallery/checkershadow`.

3. Mahmoud Afifi and Michael S. Brown. *What Else Can Fool Deep Learning? Addressing Color Constancy Errors on Deep Neural Network Performance.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 243–252, 2019.

4. Anthropic. *Claude 3.5 Sonnet*, 2024. URL: `https://www.anthropic.com/news/claude-3-5-sonnet`.

5. Michael Bach and Charlotte M. Poloschek. *Optical Illusions.* Adv Clin Neurosci Rehabil, 6(2): 20–21, 2006.

6. Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. *Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities.* arXiv preprint arXiv:2308.12966, 2023.

7. Ari Benjamin, Cheng Qiu, Ling-Qi Zhang, Konrad Kording, and Alan Stocker. *Shared Visual Illusions Between Humans and Artificial Neural Networks.* In 2019 Conference on Cognitive Computational Neuroscience, pp. 2019–1299, 2019.

8. Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. *Paligemma: A Versatile 3B VLM for Transfer.* arXiv preprint arXiv:2407.07726, 2024.

9. Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. *Breaking Common Sense: Whoops! A Vision-and-Language Benchmark of Synthetic and Compositional Images.* In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2616–2627, 2023.

10. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. *Language Models Are Few-Shot Learners.* Advances in Neural Information Processing Systems, 33: 1877–1901, 2020.

11. Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. *Sparks of Artificial General Intelligence: Early Experiments with GPT-4.* arXiv preprint arXiv:2303.12712, 2023.

12. Claus-Christian Carbon. *Understanding Human Perception by Human-Made Illusions.* Frontiers in Human Neuroscience, 8: 566, 2014.

13. Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. *How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites.* arXiv preprint arXiv:2404.16821, 2024.

14. Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. *Comparison of Deep Neural Networks to Spatio-Temporal Cortical Dynamics of Human Visual Object Recognition Reveals Hierarchical Correspondence.* Scientific Reports, 6(1): 27755, 2016.

15. Bevil R. Conway, Akiyoshi Kitaoka, Arash Yazdanbakhsh, Christopher C. Pack, and Margaret S. Livingstone. *Neural Basis for a Powerful Static Motion Illusion.* Journal of Neuroscience, 25(23): 5651–5656, 2005. doi: 10.1523/JNEUROSCI.1084-05.2005. URL: `https://www.jneurosci.org/content/25/23/5`

16. Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven Hoi. *InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning.* Advances in Neural Information Processing Systems, 36, 2024.

17. Michael R.W. Dawson. *Cognitive Impenetrability.* Springer International Publishing, Cham, 2017. doi: 10.1007/978-3-319-47829-6_1596-1. URL: `https://doi.org/10.1007/978-3-319-47829-6_1596-1`.

18. Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. *Seeing It All: Convolutional Network Layers Map the Function of the Human Visual System.* NeuroImage, 152: 184–194, 2017.

19. Jinyu Fan and Yi Zeng. *Challenging Deep Learning Models with Image Distortion Based on the Abutting Grating Illusion.* Patterns, 4(3), 2023.

20. James Fraser. *A New Visual Illusion of Direction.* British Journal of Psychology, 2(3): 307, 1908.

21. Maurizio Gentilucci, Sergio Chieffi, Elena Daprati, M. Cristina Saetti, and Ivan Toni. *Visual Illusion and Action.* Neuropsychologia, 34(5): 369–376, 1996.

22. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, 2014.

23. Alex Gomez-Villa, Adrián Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo. *On the Synthesis of Visual Illusions Using Deep Generative Models.* Journal of Vision, 22(8): 2–2, 2022.

24. Alexander Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, and Marcelo Bertalmío. *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.

25. Alexander Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo. *Color Illusions Also Deceive CNNs for Low-Level Vision Tasks: Analysis and Implications.* Vision Research, 176: 156–174, 2020.

26. Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

27. Richard L. Gregory. Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1121–1127, 1997a.

28. Richard L. Gregory. Visual illusions classified. *Trends in Cognitive Sciences*, 1(5):190–194, 1997b.

29. Richard Langton Gregory. Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 171(1024):279–296, 1968.

30. Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz Grand Challenge: Answering visual questions from blind people. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

31. Uri Hasson, Talma Hendler, Dafna Ben Bashat, and Rafael Malach. Vase or face? A neural correlate of shape-selective grouping processes in the human brain. *Journal of Cognitive Neuroscience*, 13(6):744–753, 2001.

32. Hermann von Helmholtz. Concerning the perceptions in general, 1867. *Readings in the History of Psychology*, 1948.

33. Elad Hirsch and Ayellet Tal. Color visual illusions: A statistics-based computational model. *Advances in Neural Information Processing Systems*, 33:9447–9458, 2020.

34. Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. *European Conference on Computer Vision*, pp. 727–739, Springer, 2016.

35. Charles H. Judd. The Müller-Lyer illusion. *The Psychological Review: Monograph Supplements*, 1905.

36. Daniel Kahneman. *Thinking, Fast and Slow*. Macmillan, 2011.

37. Gaetano Kanizsa. Subjective contours. *Scientific American*, 234(4):48–53, 1976.

38. Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024.

39. Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: You see what you think? Or you think what you see? *arXiv preprint arXiv:2310.14566*, 2023a.

40. Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023b.

41. Ben Lonnqvist, Alban Bornet, Adrien Doerig, and Michael H. Herzog. A comparative biology approach to DNN modeling of vision: A focus on differences, not similarities. *Journal of Vision*, 21(10):17–17, 2021.

42. OpenAI. Vision - OpenAI API, 2023. URL: `https://platform.openai.com/docs/guides/vision`.

43. OpenAI. GPT-4o, 2024. URL: `https://openai.com/index/hello-gpt-4o/`.

44. Lionel S. Penrose and Roger Penrose. Impossible objects: A special type of visual illusion. *British Journal of Psychology*, 1958.

45. Baingio Pinna, Gavin Brelstaff, and Lothar Spillmann. Surface color from boundaries: A new ' watercolor ' illusion. *Vision Research*, 41(20):2669–2676, 2001.

46. 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*, pp. 8821–8831. PMLR, 2021.

47. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

48. Ruth Rosenholtz. What your visual system sees where you are not looking. In *Human Vision and Electronic Imaging XVI*, volume 7865, pp. 343–356. SPIE, 2011.

49. R.N. Shepard. *Mind Sights: Original Visual Illusions, Ambiguities, and Other Anomalies, with a Commentary on the Play of Mind in Perception and Art*. W.H. Freeman and Company, 1990. ISBN 9780716721345. URL: `https://books.google.com.bd/books?id=1vtlQgAACAAJ`.

50. Eric D. Sun and Ron Dekel. ImageNet-trained deep neural networks exhibit illusion-like response to the scintillating grid. *Journal of Vision*, 21(11):15–15, 2021.

51. Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023.

52. Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

53. Nicholas J. Wade and Patrick Hughes. Fooling the eyes: Trompe-l ' oeil and reverse perspective. *Perception*, 28(9):1115–1119, 1999.

54. Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. GPT-4V(ision) for robotics: Multimodal task planning from human demonstration. *arXiv preprint arXiv:2311.12015*, 2023. URL: `https://api.semanticscholar.org/CorpusID:265295011`.

55. Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual expert for pretrained language models, 2023.

56. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

57. Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

58. Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

59. Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

60. Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5718–5728, 2023.

61. Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

62. Friedrich Zöllner. Ueber eine neue Art von Pseudoskopie und ihre Beziehungen zu den von Plateau und Oppel beschriebenen Bewegungsphänomenen. *Annalen der Physik*, 186(7):500–523, 1860.

63. Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6748–6758, 2023.

Supplementary Material: Appendices

A Preliminaries

本セクションでは、光学的錯覚に関する認知科学の関連文献を概観します。

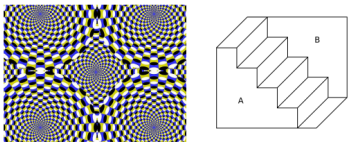


Figure 6: Left: Kitaoka's 'Throwing cast nets' (Conway et al., 2005), an example of cognitive impenetrability. Even though we know it is a static image, the illusion of motion persists. Right: Schroeder stairs (Rosenholtz, 2011), an example of an ambiguous or bistable image. The drawing may be perceived as either a staircase leading from left to right downward or as the same staircase only turned upside down.

Optical Illusion（光学的錯覚）

光学的錯覚とは、視覚システムが現実を誤って解釈し、知覚されるものと実際に存在するものとの間に不一致が生じる視覚知覚現象を指します。研究者たちは、光学的錯覚は主に人間の知覚的知識と概念的知識との間の不一致から生じると提案しています（Gregory, 1997b; Carbon, 2014）。具体的な例として、私たちの速い知覚的知識は図**??**を象として知覚しますが、注意深く検討すると、この図が幾何学に関する私たちの既存の概念と一致しないことが理解できます。

Cognitive Impenetrability（認知不可侵性）

認知不可侵性とは、基本的な認知情報処理に関する特性を指します。あるプロセスが認知不可侵性を持つ場合、その操作は、エージェントの信念、欲求、目標などの心的表象の内容に変化が生じても影響を受けません。これは、このプロセスが「配線済み」であり、神経科学を通じてのみ説明できるためです（Dawson, 2017）。図**??**に示される視覚的錯覚は、この現象を示しており、錯覚の存在を認識している場合でも錯覚は持続します。

Ambiguous or Bi-stable Images（曖昧または双安定画像）

曖昧または双安定画像とは、2 つ以上の明確に異なる形態の間で曖昧さを生み出す画像を指します。このような画像の古典的な例には、ルビンの壺（Hasson et al., 2001）、シュレーダーの階段（図**??**）、およびネッカーの立方体（Rosenholtz, 2011）などがあります。

B Object Localization with VLMs（VLM を用いた物体の位置特定）

物体検出（Object Detection）（Girshick et al., 2014）は、コンピュータビジョンの中心的なタスクであり、モデルが画像内の物体のバウンディングボックスの座標を出力します。純粋なビジョンモデル、例えば DETR（Zong et al., 2023）は物体検出において最先端の性能を示していますが、事前学習された VLM（Wang et al., 2023; Bai et al., 2023）を使用した物体検出にも大きな進展が見られます。しかし、多くの VLM はネイティブに物体検出をサポートしておらず、視覚専用モデルよりも著しく性能が劣ります。

そのため、本研究では簡易なタスクであるソフトローカリゼーション（soft localization）を選択しました。このタスクでは、VLM に物体の大まかな位置（例えば、左側または右側）を出力するようにプロンプトを与え、正確なバウンディングボックス座標を求めることはしません。

C Literature Review of Illusion Classification（錯覚分類の文献レビュー）

光学的錯覚（optical illusions）の分類は科学界で大きな課題となっており、決定的な合意には至っていません。本研究では、IllusionVQA の分類に先立つものとして、Gregory（1997b）および Bach Poloschek（2006）が提案した 2 つの主要な分類アプローチをレビューします。

Gregory（1997b）は、光学的錯覚をその根本的な原因に基づき 4 つに分類しました：

- **物理的錯覚（Physical illusions）**：霧やレンズの歪みによって引き起こされるような、画像形成を直接的に妨害する現象。
- **生理学的錯覚（Physiological illusions）**：目に圧力をかけたときに発生するきらめきなど、視覚システム内の障害によって生じる現象。
- **知識ベースの錯覚（Knowledge-based illusions）**：2 次元のスケッチを 3 次元の物体として解釈するような、感覚入力を誤って解釈するための先行知識の活用。
- **ルールベースの錯覚（Rule-based illusions）**：カニッツァ三角形（Kanizsa, 1976）のように、視覚システムがパターンを補完したり、欠落した情報を推測する傾向を利用することで、錯覚的な知覚を引き起こす現象。

D IllusionVQA-Comprehension Categories and Question Types

Class	#	Description	Example
Impossible Object	134	Objects that cannot exist in 3D space.	Penrose Triangle (Penrose & Penrose, 1958)
Real-Scene	64	Hard to interpret due to forced perspective, unnatural poses, objects being obscured, etc.	
Size	46	Two objects with the same dimensions look different or vice versa.	Müller-Lyer illusion (Judd, 1905)
Hidden	45	(Silhouettes of) certain objects are present somewhere without being immediately obvious.	
Deceptive Design	37	Items have been painted to appear as something else.	Trompe-l'œil art (Wade & Hughes, 1999)
Angle Illusion	26	Parallel lines look curved or angled, and vice versa.	Zöllner illusion (Zöllner, 1860).
Color	23	Different colors or shades appear the same, and vice versa.	Checker Shadow Illusion (Adelson, 1995)
Edited-Scene	21	Edited to depict something impossible although it may appear normal on a cursory view.	
Upside-Down	7	Two intertwined entities depending on vertical orientation.	
Positive-Negative Space	7	Images that create ambiguity between two or more distinct forms represented by two or more colors.	Rubin's vase (Hasson et al., 2001)
Circle-Spiral	6	Circles appear as spirals due to peculiar color or background pattern.	Fraser Spiral (Fraser, 1908)
Miscellaneous	19		
Total	435		

Table 4: Types of Illusions in IllusionVQA-Comprehension - Test Split

Major Question Types	Count
"What is unusual about this image?"	114
"Describe this image."	80
"What is hidden in this image?"	22
Image-specific Questions	219

Table 5: Question Types in IllusionVQA-Comprehension - Test Split.

さらに、Gregory（1997b）は自然言語に根ざした外観に基づく光学的錯覚の補完的な分類を提案しています。この分類には以下が含まれます：

- **曖昧性（Ambiguities）**（図 6）
- **歪み（Distortions）**（図 2b）
- **逆説（Paradoxes）**（図 2a）
- **虚構（Fictions）**（図 2c）

本研究では、セクション 3.3 において、IllusionVQA に対してこの分類スキームを採用しています。

E Few-shot and Chain-of-Thought Evaluation Prompts（Few-shot および Chain-of-Thought 評価のプロンプト）

0-shot Instruction（0 ショット指示）以下の指示に従ってください。画像、指示、および選択肢が与えられます。正しい選択肢を選び、対応する文字のみを答えてください。理由を説明してはいけません。解答を繰り返してはいけません。

> You'll be given an image, an instruction, and some options. You have to select the correct one. Do not explain your reasoning. Answer with only the letter that corresponds to the correct option. Do not repeat the entire answer.

4-shot Instruction（4 ショット指示）以下の指示に従ってください。画像、指示、および選択肢が与えられます。正しい選択肢を選び、対応する文字のみを答えてください。理由を説明してはいけません。解答を繰り返してはいけません。

> You'll be given an image, an instruction, and some options. You have to select the correct one. Do not explain your reasoning. Answer with only the letter that corresponds to the correct option. Do not repeat the entire answer. Here are a few examples: {few-shot examples} Now you try it.

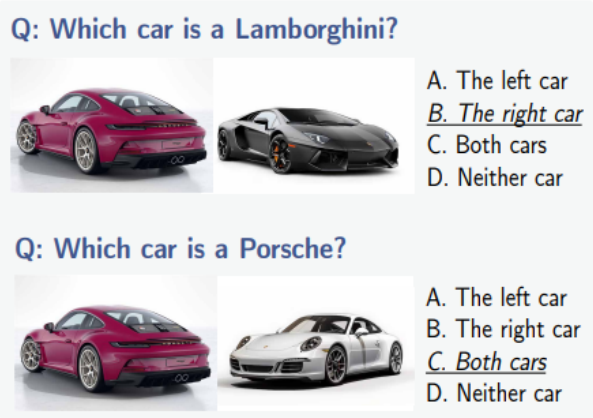


Figure 7: Soft localization of ordinary objects.

VLM	Prompt Type	Accuracy
InstructBLIP	0-shot	30
LLaVA-1.5	0-shot	29.25
CogVLM	0-shot	49
Gemini-Pro	0-shot	100
GPT4V	0-shot	100

Table 6: Accuracy of localizing ordinary objects.

4-shot+CoT Instruction（4 ショット+Chain-of-Thought 指示）以下の指示に従ってください。画像、指示、および選択肢が与えられます。正しい選択肢を選ぶ際に、質問および画像の文脈に基づいて選択肢を推論してください。解答を「Answer: {letter_of_correct_choice}」で終えてください（波括弧は含まないでください）。以下にいくつかの例があります：{few-shot examples} Now you try it.

> You'll be given an image, an instruction, and some choices. You have to select the correct one. Reason about the choices in the context of the question and the image. End your answer with 'Answer: {letter_of_correct_choice}' without the curly brackets. Here are a few examples: {few-shot examples} Now you try it.

CoT Reasoning Template（Chain-of-Thought 推論テンプレート）

> The left object is a {description} which is {possible/impossible}. The right object is a {description} which is {possible/impossible}.

F Soft Localization with Ordinary Objects（通常の物体を用いたソフトローカリゼーション）

IllusionVQA の Soft Localization タスクにおける VLM の性能が予想以上に低かったため、VLM の通常の物体を位置特定する能力をテストするための簡易補助データセットを作成しました。具体的には、異なるブランドのスポーツカーの画像を使用し、モデルに特定のブランドを特定するよう求めます。2 つの物体が視覚的に似ているため、このタスクは中程度の難易度を維持しています。

Figure 7: この新しいデータセットのサンプルを示しています。このタスクのために 400 のインスタンスを収集し、すべての VLM を 0 ショット設定でテストしました。

Table 6: 小型 VLM はこのタスクでの性能が低い一方で、GPT4V と Gemini-Pro は完全な正確性を達成しました。この結果は、大型 VLM がシーン内で物体を特定できることを確認するものです。

G Human Evaluation（人間の評価）

本研究では、筆者の所属機関の 3 人の非著者である学部生に協力を依頼し、光学的錯覚の理解および位置特定タスクにおける人間の性能を評価しました。この評価はインターネットアクセスなしで実施され、評価者には事前にタスクが光学的錯覚に関連していることを通知しませんでした。各評価者には以下の指示が与えられた WebUI（図 8 参照）を通じて評価を行いました。

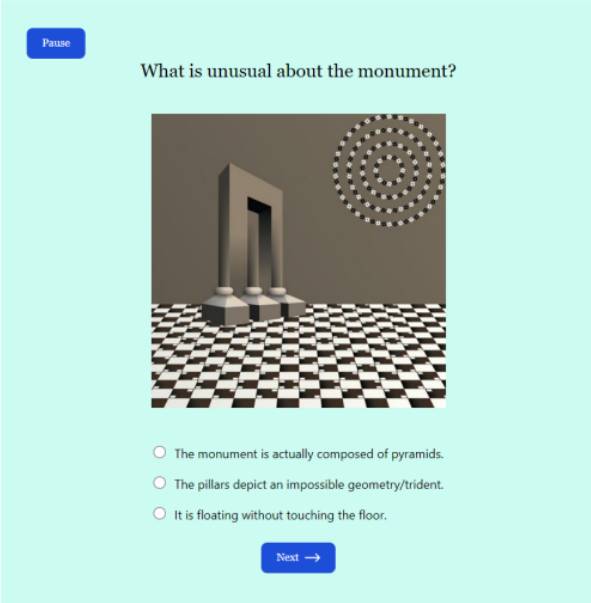


Figure 8: WebUI presented to human evaluators.

Evaluator	Accuracy	Avg. Response Time
1	87.13	16.46
2	86.67	12.95
3	92.87	15.57

Table 7: Human performance and response time on IllusionVQA-Comprehension.

評価者への指示: 「各ページには、光学的錯覚に関する質問と画像、続いていくつかの選択肢が表示されます。質問に最も適した選択肢を選んでください。錯覚は誤解を招く可能性があるため、画像内の実際の対象がどのようであるかを反映した選択肢を選んでください。必要に応じてズームインまたはズームアウトして、より良い視点を得てください。各回答は内部的にタイム計測されていますので、必要に応じて「一時停止」ボタンを使用してください。」

すべての評価者には同じ順序で質問が提示されました。表 7 に示すように、評価者個々の正答率は比較可能な結果を示しています。評価者間のペアワイズ Cohen のカッパ係数はそれぞれ 0.808、0.796、0.773 であり、評価者間に実質的な一致があることを示しています[3]。

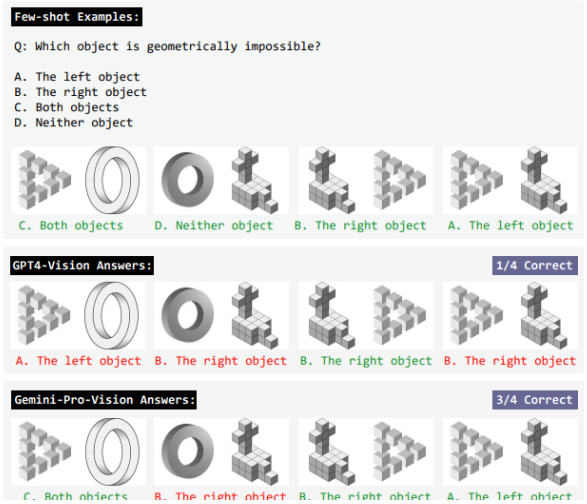
IllusionVQA-Soft-Localization タスクにおいて、すべての評価者が 200 個の選択されたインスタンスで完全な正確性を達成しました。平均応答時間は 5.68 秒でした。

H Limits of In-Context Learning For Optical Illusions（光学的錯覚におけるインコンテキスト学習の限界）

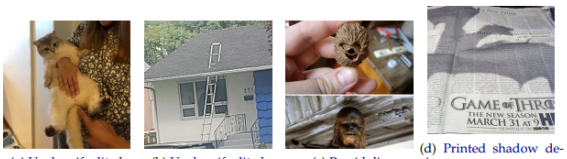
図 9: GPT4V および Gemini-Pro は、幾何学的に不可能なオブジェクトの位置を特定するタスクにおいて、一貫して失敗しました。この現象は、正しい回答が文脈内に提示されている場合（4-shot 設定）でも確認されました。

表 8: IllusionVQA-Comprehension の結果（2024 年 7 月時点での最新の VLM を含む）。InternVLM2（Chen ら、2024）、PaliGemma-3B（Beyer ら、2024）、GPT-4o（OpenAI、2024）、および Claude 3.5 Sonnet Anthropic（2024）の評価を追加しました。4-shot プロンプトを使用することで精度が向上したカテゴリは下線で示されています。

[3]Cohen のカッパ係数の解釈については関連文献を参照してください。



I Examples of Images Removed During Preprocessing



(a) Unclear if edited (b) Unclear if edited (c) Pareidolia (d) Printed shadow design

Figure 10: Examples of images we filtered out during pre-processing. These images were found in online archives of optical illusions and are colloquially considered illusions.

J Updated IllusionVQA-Comprehension Results

Class	#	0-shot						4-shot				Human
		InternVL2 8B	Pali Gemma	Gemini 1.0 Pro	Claude 3.5 Sonnet	GPT4V	GPT4o	Gemini 1.0 Pro	Claude 3.5 Sonnet	GPT4V	GPT4o	
Impossible Object	134	49.25	32.09	56.72	**64.93**	55.22	63.43	56.72	63.43	58.96	61.94	98.51
Real-Scene	64	40.63	35.94	46.88	54.69	57.81	**64.06**	46.88	57.81	54.69	57.81	98.44
Size	46	43.48	15.22	45.65	50.00	**58.70**	45.65	52.17	80.43	69.57	93.47	63.04
Hidden	45	44.44	33.33	42.22	37.78	51.11	**66.67**	48.89	44.44	46.67	48.89	100
Deceptive Design	37	37.84	32.43	64.86	70.27	70.27	**72.97**	67.57	67.57	72.97	78.38	94.59
Angle Illusion	26	50.00	26.92	53.85	**73.08**	69.23	50.00	50.00	73.08	84.62	80.77	84.62
Color	23	26.09	34.78	17.39	65.22	**69.57**	52.17	17.39	86.96	82.61	78.26	60.87
Edited-Scene	21	66.67	42.86	66.67	61.90	71.43	**80.95**	66.67	71.43	80.95	85.71	100
Upside-Down	7	**85.71**	42.86	57.14	**85.71**	71.43	71.43	57.14	85.71	71.43	42.86	100
Pos.-Neg. Space	7	42.86	42.86	**85.71**	71.43	57.14	**85.71**	71.43	71.43	85.71	71.43	100
Circle-Spiral	6	0.00	0.00	33.33	33.33	**50.00**	**50.00**	33.33	50.00	33.33	50.00	66.67
Miscellaneous	19	42.11	31.58	**52.63**	47.37	42.11	**52.63**	57.89	52.63	42.11	52.63	89.47
Total	435	45.06	31.26	51.26	59.08	58.85	**62.53**	52.87	66.44	62.99	67.12	91.03

Table 8: IllusionVQA-Comprehension results updated to include state-of-the-art VLMs as of July 2024. Added evaluation of InternVLM2 (Chen et al., 2024), PaliGemma-3B (Beyer et al., 2024), GPT-4o (OpenAI, 2024) and Claude 3.5 Sonnet Anthropic (2024). Categories where accuracy has improved using 4-shot prompting are underlined.