大規模視覚言語モデルは錯視を理解しているか

篠崎 大河 [2,1]　土井 智暉 [1]　綿引 周 [1]　西田 知史 [3,4,5]　谷中 瞳 [1]
[1] 東京大学　[2] 慶應義塾大学　[3] 情報通信研究機構　[4] 大阪大学　[5] 北海道大学
snzktig@keio.jp　amanew@g.ecc.u-tokyo.ac.jp　s-nishida@nict.go.jp
{doi-tomoki701, hyanaka}@is.s.u-tokyo.ac.jp

概要

錯視画像とは，その実際の特徴と見かけの特徴が異なるような画像のことである．大規模視覚言語モデル (Large Vision-Language Model: LVLM) に関して，その錯視画像の認識能力を評価する研究が近年行われている．先行研究においては，実験の結果により，LVLM は錯視に騙されやすい，あるいは人間と同様の騙され方をすると考えられている．しかし，先行研究は錯視に関する重要な区別を見落としているため，その実験結果に曖昧さを残している．本研究では，可能な限り曖昧さを排した手法を提案し，それを用いて LVLM の錯視認識能力を評価する．実験の結果，LVLM は一見して錯視を理解しているように思われるものの，実際には錯視に関する一般的な知識から回答しており，錯視を視覚的に正しく認識しているわけではないことが示唆される．

1　はじめに

錯視画像は，その実際の特徴と見かけの特徴が異なるような画像である [1]．これまで，深層ニューラルネットワークを対象としてその錯視画像の認識能力を評価する研究が行われており，それらは人間と同様の騙され方をすると報告されている [2, 3, 4, 5, 6]．これに続いて近年，大規模視覚言語モデル (Large Vision-Language Model: LVLM) に関して，同様の評価の試みがある [7, 8, 9]．これらの研究の背後には，LVLM が錯視に騙されるのか否か，そして，錯視に騙される人間の指示を正しく理解する能力を持つかという問いに答えたいという動機がある [9]．例えば，図 1 の A について，人間が「大きい方の赤い円」と述べたときに，どちらの円を指しているか理解できるか，といったことである．

先行研究では，一方で，GPT-4V，Gemini Pro Vision，Claude 3，LLaVA-1.5 などの LVLM は錯視に騙されやすいことや [7]，LLaVA-1.5 や BLIP などのモデルで，モデルサイズが大きいほど人間と同様の騙され方をすることが報告されている [9]．他方で，LVLM は錯視に騙されやすいものの，few-shot prompting を用いると騙されにくくなるという報告もある [8]．

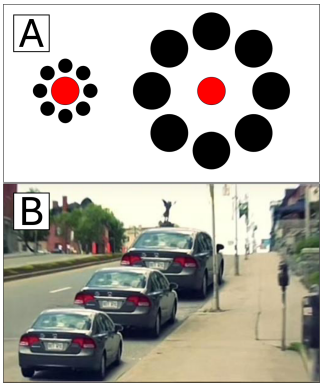


図 1　A: 抽象的な錯視画像．左の赤い円が大きく見えるが，二つの赤い円は同じ大きさである．B: 非抽象的な錯視画像．最も奥にある車が最も大きく見えるが，三台の車の幅と高さは画像上の同じ距離を占めている [8]．

しかし，先行研究の実験設定は錯視認識能力の正確な評価において 2 つの問題があり，実験結果に曖昧さが残る．

第一に，先行研究は抽象的画像だけでなく非抽象的画像を評価対象の錯視に含めている．抽象的画像が二次元の抽象的な図形を表象するものであるのに対し，非抽象的画像は三次元空間中の対象を表象するものである（図 1）．非抽象的画像に関する質問は常に二通りの仕方で解釈されうる．例えば，図 1 の B に関して，「どの車が最も大きいか」という質問について考えよう．もしこの質問が画像そのものについてのものであると解釈されたなら，正解は「全ての車は同じ大きさである」というものになる．実際，ある先行研究はこれを正解としている [8]．しかし，もし質問が画像によって表象されている対象，すなわち，現実の状況における車についてのものであると解釈されたなら，正解は「カメラから最も遠い車が最も大きい」というものになる．なぜな

ら，現実の状況において，観察者から遠い対象が近い対象と視野の同じ面積を占めている場合，遠い対象の方が大きいはずであるからである．それゆえ，いずれの回答も正解と解釈されうる．したがって，LVLM がどちらを答えたとしても，その錯視の認識能力を評価する上では曖昧さが残る．

第二に，先行研究は画像の実際の特徴と見かけの特徴を区別していない [7, 8, 9]．実際の特徴とは画像が実際に持つ特徴のことであり，逆に，見かけの特徴とは画像が持つように見える特徴のことである．通常の画像において見かけの特徴は実際のそれと一致している．しかし，錯視画像においては両者は異なる．例えば，図 1 の A において，左の赤い円は右の赤い円と同じ大きさである．それゆえ，「右の赤い円と同じ大きさである」は左の円が持つ実際の特徴である．しかし，左の円は右より大きく見える．それゆえ，「右の赤い円より大きい」は左の円が持つ見かけの特徴である．したがって，この区別を明示せずに LVLM に質問することはさらなる曖昧さを生じさせる．というのも，どちらの特徴に LVLM が着目しているかが明示されないからである．

本研究では，可能な限り曖昧さを排した手法を提案し，それを用いて LVLM の錯視認識能力を評価する．具体的には，i) 抽象的画像のみのデータセットを構築し，かつ，ii) 実際の特徴と見かけの特徴を区別するプロンプトを用いた上で，通常の錯視画像だけでなく，見かけの特徴を改変した偽の錯視画像の認識を評価する．結果として，LVLM は一見して錯視を理解しているように思われるものの，実際には錯視に関する一般的な知識から回答しており，錯視を視覚的に正しく認識しているわけではないことが示唆される．また，本データセットは研究利用可能な形で公開予定である．

2 提案手法

2.1 タスク

本研究では，LVLM の錯視の認識能力を評価するため，i) 本物の錯視画像の特徴を答えさせるタスク（真正錯視画像質問応答：真正錯視 VQA），ii) 偽の錯視画像の特徴を答えさせるタスク（偽錯視 VQA），iii) それぞれの錯視画像から錯視を引き起こす要素を取り除いた画像の特徴を答えさせるタスク（統制 VQA）を作成した．また，画像の実際の特徴と見かけの特徴を区別しないことによる曖昧さを回避するため，それぞれの特徴について問う二つの質問セットを作成し，それぞれにおいてどちらの特徴について問うているかを明示した．実際にモデルの評価に使用したプロンプトは付録 A に示す．

真正錯視 VQA　においては，錯視の効果によって実際の特徴と見かけの特徴が異なる画像（すなわち通常の錯視画像）を提示し，それら 2 種類の特徴をそれぞれ問う（表 1）．

偽錯視 VQA　においては，錯視の効果を持つものの，その見かけの特徴に合うように実際の特徴を修正した偽の錯視画像を提示し，真正錯視 VQA と同じ質問を問う．例えば，表 1 で示す偽錯視画像において，左の赤い円は右の赤い円より大きく見え，かつ，実際にそれは右の円より大きい．それゆえ，偽錯視 VQA においては実際の特徴に関する質問の正解のみ真正錯視 VQA と異なる（表 1 の赤字部分）．もしある LVLM が錯視を視覚的に正しく認識しているならば，偽錯視 VQA の実際の特徴について正答できるはずである．

統制 VQA　においては，真正錯視 VQA と偽錯視 VQA で用いた画像から錯視を引き起こす要素を取り除いた画像（付録 B）を提示し，それぞれと同じ質問を問う．本研究ではそれらの画像を，真正錯視画像あるいは偽錯視画像と対応する統制画像と呼ぶ．これは真正錯視 VQA と偽錯視 VQA との結果との比較用である．

2.2 データセット

本研究におけるデータセットの作成は，i) 錯視の選定，ii) 画像の作成，iii) 画像の人手評価の 3 ステップで構成される．

錯視の選定　本研究の目的に照らして，以下の条件を満たす錯視を，有名な錯視の中から選定した．

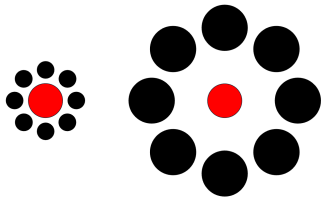
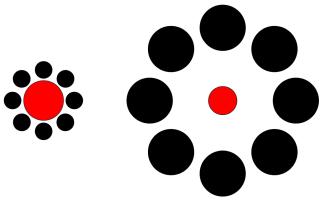
- 見かけの特徴と実際の特徴が異なるというタイプの錯視であること（錯視タイプの統一のため）[1)]
- 抽象的画像であること（非抽象的画像に伴う曖昧さ回避のため）

このステップで選定した錯視は 15 種である．

画像の作成　15 種の錯視それぞれについて，角度や色など錯視を引き起こす要素と無関係な要素を変化させた真正錯視画像を 2 枚ずつ作成した（合計 30 枚）．また，それぞれの真正錯視画像と対応する偽

1) 先行研究ではそれ以外のタイプの画像（不可能立体の絵など）が「錯視」として扱われることもある [8].

表 1　錯視画像と質問

タスク	真正錯視 VQA	偽錯視 VQA
画像		
錯視名	エビングハウス錯視	
質問（実際の特徴）	Which red circle is bigger?	
選択肢（実際の特徴）	“The left red circle is bigger.”, “The right red circle is bigger.”, “Both red circles are the same size.”	
正解（実際の特徴）	Both red circles are the same size.	The left red circle is bigger.
質問（見かけの特徴）	Which red circle appears bigger?	
選択肢（見かけの特徴）	“The left red circle appears bigger.”, “The right red circle appears bigger.”, “Both red circles appear the same size.”	
正解（見かけの特徴）	The left red circle appears bigger.	

錯視画像（合計 30 枚）を作成した．さらに，それぞれの真正錯視画像および偽錯視画像と対応する統制画像（合計 60 枚）を作成した．真正錯視画像は，多くの人間から見て，実際の特徴と異なる見かけの特徴を持つことを意図して作成されている．偽錯視画像は，多くの人間から見て，両者が異ならず，かつ，実際の特徴を正しく推測できるように意図して作成されている．

画像の人手評価　我々が意図した特徴を，作成した画像が実際に持つかを確かめるため，作成した画像を代表する真正錯視画像 17 枚，偽錯視画像 17 枚，統制画像 34 枚の合計 68 枚を用いて，52 名の人間を対象にそれらの見かけの特徴と実際の特徴を問う実験を行った．評価においては，各設問の最頻回答を被験者を代表する回答とした．結果として，代表的な回答は，真正錯視画像 1 枚を除き，すべての画像について我々が意図した特徴を持つと認めるものであった．意図した特徴を認める回答が得られなかった錯視を表現する 2 枚[2]の錯視画像および，それと対応する偽錯視画像と統制画像はデータセットから除外した．最終的に 14 種の錯視を使用し，真正錯視画像 28 枚，偽錯視画像 28 枚，それぞれに対応する統制画像 56 枚をデータセットに含めた．使用した錯視の一覧は付録 E に示す．

2）　1 種の錯視につき 2 枚の画像を作成したためである．

3　実験 1：真正錯視 VQA

3.1　実験設定

評価の対象とするモデルは GPT-4o [3] , Claude3.5 [4] , LLaVA-NeXT（72b, 110b）[10, 11] である．それらに真正錯視画像セットを与え，それぞれの画像についてその実際の特徴と見かけの特徴を zero-shot prompting で回答させた．

3.2　結果と分析

真正錯視 VQA における各モデルの回答を i) 実際の特徴と見かけの特徴の両方の質問に正答，ii) 見かけの特徴のみに正答，iii) 実際の特徴のみに正答，iv) 両方誤答のいずれかに分類し，それぞれの種類の回答の割合を算出した（表 2）．統制画像に対する回答との比較は，最も正答率の高い GPT-4o のもののみ付録 C に示す．結果として，LLaVA-NeXT はサイズにかかわらずほとんどの設問で見かけの特徴か実際の特徴のいずれかに関して誤答するのに対し，GPT-4o と Claude3.5 は見かけの特徴と実際の特徴の両方に関して正答率が高いため，一見して錯視を理解しているかのように思われる．

3）　https://openai.com/index/GPT-4o-system-card/
4）　https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf

モデル	GPT-4o	Claude3.5	LLaVA-72b	LLaVA-110b	Humans
両方正答	75.0	60.7	0.0	7.1	62.5
見かけのみ	7.1	7.1	28.6	25.0	37.5
実際のみ	14.3	32.1	35.7	42.9	0.0
両方誤答	3.6	0.0	35.7	25.0	0.0

表 2　真正錯視 VQA における正答率. 見かけの特徴に関して人間の誤答がないのは, 我々が意図した見かけの特徴を認める回答が得られなかった画像をデータセットから除外したためである（第 2.2 節参照）.

モデル	GPT-4o	Humans
両方正答	17.9	100.0
見かけのみ	75.0	0.0
実際のみ	0.0	0.0
両方誤答	7.1	0.0

表 3　偽錯視 VQA における正答率.

4　実験 2：偽錯視 VQA

4.1　実験設定

評価の対象とするモデルは真正錯視 VQA と同じである．これに偽錯視画像セットを与え，それぞれの画像についてその実際の特徴と見かけの特徴を zero-shot prompting で回答させた.

4.2　結果と分析

偽錯視 VQA における各モデルの回答を真正錯視 VQA と同様に分類した．各分類の回答の割合の算出においては，それぞれのモデルが実際の特徴と見かけの特徴に関して両方正答した真正錯視画像に対応する偽錯視画像に対する回答のみ計上した．ここでは，真正錯視 VQA において最も成績の高かった GPT-4o の結果を示す（表 3）． 統制画像に対する回答との比較は付録 C に示す.

表 2 の真正錯視 VQA の結果と異なり，見かけの特徴のみについて正答する傾向が見られた．これは，図 2 に示すような回答パターンが多いことを意味する． なお，Claude3.5 においても同様の結果が得られた（付録 D）．真正錯視 VQA において両方正答した画像と対応する偽錯視画像についてこのパターンが多いということは，LVLM が画像の見かけの特徴を視覚的に正しく理解しているのではなく，錯視に関する一般的な知識からその特徴を推測しているだけであることを示唆する．例えば，図 2 の設問に関して言えば，偽錯視画像を真正錯視画像（エビングハウス錯視画像）と誤って判断し，「エビングハウス錯視において二つの円の実際の大きさは同じである」という一般的な知識から，偽錯視画像における二つの赤い円の実際の大きさが同じであると判断するということである．紙幅の関係で掲載することが叶わないが，他のプロンプト[5)]でも同様の結果が得られた．以上より，LVLM が錯視を視覚的に正しく理解しているわけではないことが示唆される.

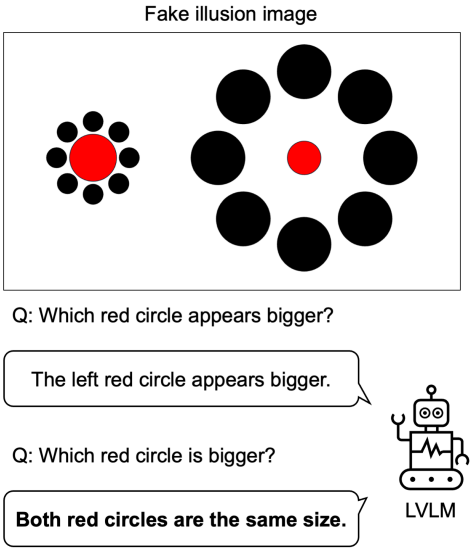


図 2　偽錯視画像タスクのモデルの回答例．偽錯視画像について，見かけの特徴についての質問には正答するが実際の特徴についての質問には誤答する（太字強調部分）.

5　おわりに

本研究では，LVLM の錯視の認識能力の評価に関して，先行研究の欠点を補う手法として、真正錯視と偽錯視の両方について見かけと実際の特徴を問うタスクを提案し，実施した．真正錯視 VQA の結果から，LVLM は一見して錯視を理解しているように思われるものの，偽錯視 VQA の結果から，実際には錯視に関する一般的な知識から回答しており，錯視を視覚的に正しく認識しているわけではないと示唆される．今後，画像の種類を増やしてさらに調査する予定である.

5)　few-shot prompting およびメタ認知を促すプロンプト [12].

謝辞

本研究は JSPS 科研費学術変革領域研究（B）「ナラティブ意識学」JP24H00809 の支援を受けたものである.

参考文献

[1] RH Day. The nature of, perceptual illusions. **Interdisciplinary Science Reviews**, Vol. 9, No. 1, pp. 47–58, 1984.

[2] 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.

[3] 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**, Vol. 10, pp. 2019–1299. Cognitive Computational Neuroscience, 2019.

[4] Alexander Gomez-Villa, Adrián 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.

[5] Alexander Gomez-Villa, Adrián Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo. Color illusions also deceive cnns for low-level vision tasks: Analysis and implications. **Vision Research**, Vol. 176, pp. 156–174, 2020.

[6] Eric D Sun and Ron Dekel. Imagenet-trained deep neural networks exhibit illusion-like response to the scintillating grid. **Journal of Vision**, Vol. 21, No. 11, pp. 15–15, 2021.

[7] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 14375–14385, June 2024.

[8] Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models. **Computing Research Repository**, Vol. arXiv:2403.15952, , 2024. version 3.

[9] Yichi Zhang, Jiayi Pan, Yuchen Zhou, Rui Pan, and Joyce Chai. Grounding visual illusions in language: Do vision-language models perceive illusions like humans?, 2023.

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[11] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024.

[12] Yuqing Wang and Yun Zhao. Metacognitive prompting improves understanding in large language models, 2024.

A　プロンプト

実際の特徴を問うプロンプト

You will be asked to answer a question about the actual feature of the figure. This question asks you what features the figure actually has. I will provide answer options. Choose one of the options to answer the question by guessing the actual features of the figure, regardless of how it appears subjectively to you.

見かけの特徴を問うプロンプト

You will be asked to answer a question about the apparent feature of the figure. This question asks you how the figure appears subjectively to you. I will provide answer options. Choose one of the options to answer the question as you see it, regardless of what features you think the figure actually has.

B　統制画像例

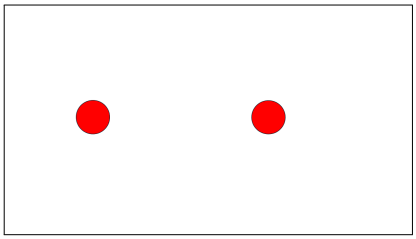


図 3　真正錯視画像に対応する統制画像の例.

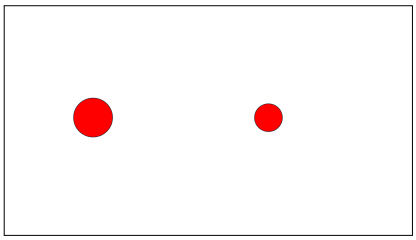


図 4　偽錯視画像に対応する統制画像の例.

C　GPT-4o の各 VQA の正答率

真正錯視 VQA における両方正答の割合はそれに対応する統制 VQA の割合と同じである（表 4）．偽錯視 VQA における両方正答の割合はそれに対応する統制 VQA の割合より低い（表 5）．このことは偽錯視 VQA における錯視の効果を持つ要素が正答率を下げたことを示唆する．ただし，真正錯視 VQA と偽錯視 VQA のそれぞれに対応する統制 VQA における両方正答の割合にも差が見られる（後者が低い）ため，錯視と無関係の要素も正答率の低下に寄与していることが示唆される．

VQA	錯視	錯視統制
両方正答	75.0	75.0
見かけのみ	7.1	0.0
実際のみ	14.3	17.9
両方誤答	3.6	7.1

表 4　GPT-4o の真正錯視 VQA とそれに対応する統制 VQA の正答率．「錯視」の列は真正錯視 VQA における各回答パターンの割合を，「錯視統制」の列は真正錯視に対応する統制 VQA 各回答パターンの割合を示している．

VQA	偽錯視	偽錯視統制
両方正答	17.9	53.6
見かけのみ	75.0	32.1
実際のみ	0.0	0.0
両方誤答	7.1	14.3

表 5　GPT-4o の偽錯視 VQA とそれに対応する統制 VQA の正答率．「偽錯視」の列は真正錯視 VQA における各回答パターンの割合を，「偽錯視統制」の列は真正錯視に対応する統制 VQA 各回答パターンの割合を示している．

D　Claude3.5 の偽錯視 VQA の正答率

モデル	Claude3.5	Humans
両方正答	7.1	100.0
見かけのみ	78.6	0.0
実際のみ	0.0	0.0
両方誤答	14.3	0.0

表 6　偽錯視 VQA における正答率.

E　使用した錯視一覧

1. カフェウォール錯視
2. ツェルナー錯視
3. ミュラーリヤー錯視
4. フィック錯視
5. ヘリング錯視
6. オービソン錯視
7. エビングハウス錯視
8. ジャストロー錯視
9. デルブーフ錯視
10. ムンカー錯視
11. 三角形分割錯視
12. ザンダー錯視
13. 彩度の恒常性を利用した錯視
14. 明度の恒常性を利用した錯視

ポンゾ錯視の画像も作成したが，2.2 節に示した理由により除外した．