視覚情報による曖昧性解消コーパスの検討

李　相明 [1,2]　河野　誠也 [2,1]　吉野　幸一郎 [3,2,1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所　ガーディアンロボットプロジェクト
[3] 東京科学大学
lee.sangmyeong.lo3@is.naist.jp　seiya.kawano@riken.jp
koichiro.yoshino@riken.jp

概要

自然言語の曖昧性の解消は、システムとユーザ間の意味の一致を達成する上で解決すべき重大な課題である。そのためには、実世界での視覚情報と自然言語の統合処理が鍵となる。本研究では、既存の Vision & Language Model を対象に、視覚情報を利用した曖昧性理解能力を評価するためのデータセットの検討を行った。さらに、試験的に収集したデータを用いて、曖昧性が解消された文と画像の識別能力を測定することで VLM が視覚情報を活用して曖昧性を解消する能力を評価した。実験結果に基づき、既存モデルの曖昧性解消能力の可能性を探求するとともに、エラーケースの分析を行った。この分析から、モデルが持つ曖昧性に対する弱点と、ベンチマークやデータセットの今後の課題について考察を行った。

1　はじめに

自然言語における曖昧性とは、一つの文が複数の解釈を持つ現象をさす。この曖昧性を解消するためには、視覚情報 [1, 2, 3, 4] や話し方の抑揚 [5] などの補助的な情報がしばしば必要となる。特に視覚情報は、実世界の理解において重要な役割を果たしており、曖昧性解消の主要な手掛かりとして注目されている。視覚情報を用いた曖昧性解消のタスクでは、文が持つ複数の意味を視覚的な文脈と結びつける必要があり、言語と視覚の両方を深く理解する高度な処理が求められる。視覚情報を活用した曖昧性解消能力は、VLM（Vision & Language Model）の実世界での応用において特に重要である。この能力により、ユーザの指示や意図の誤解を防ぐだけでなく、モデルが高度な推論を行う基盤を提供する期待される。

現在のところ、既存の VLM が視覚情報を活用して曖昧性を解消する能力についての評価は十分に検討されていない。ARO [6] や Winoground [7] などのベンチマークは、視覚情報を用いて文中の単語の関係性を理解する優れた評価ツールを提供している。しかし、これらのベンチマークで使用されている例示の多くは、曖昧性のある文を含んでいない点で課題が残る。一方で、LAVA [1] や TAB [2] のようなベンチマークは、曖昧性に特化した文と画像を結びつけるタスクを設定しているが、これらのデータセットには量や質の面での制約がある。これにより、曖昧性解消能力を正確に評価するための基盤が不十分であるといえる。

本研究は、VLM の視覚情報による曖昧性解消能力を評価し、その改善に向けたデータセットの構築を試みる。大規模視覚言語生成モデルの進展 [8, 9] により、データの収集や質の制御が容易になることが期待される。本研究では、生成モデルを用いて試験的にデータを収集し、そのデータを基に画像と曖昧性が解消された文を結びつけるタスクを設計する。これにより、既存 VLM の曖昧性への対応能力を評価し、エラーケースを分析して弱点を探る。定量評価と定性分析を通じ、データセットの不足点を明確にし、今後の研究の基盤となることを目指す。

2　先行研究

2.1　VLM：言語と視覚の表現の共有

異なるモダリティである言語と視覚を繋ぐ鍵は、同じ意味を持つ画像と文が特徴量表現を共有することにある。CLIP [10] は、画像とキャプションを対照的に学習し、それらが同じ特徴量を共有する様に設計されている。この対照学習により、多量の事前学習が可能となり、Zero 及び Few-show での識別能力を発揮し、画像生成モデルなどの基盤技術として活用されている [9, 11]。CLIP の後続研究では、文と画

表 1　本研究における曖昧性の種類

種類（略称）	例文	説明
VP attachment（VP）	The girl saw a man talking to the telephone	発話の主体は少女か、男か？
PP attachment（PP）	The boy approached the chair with a bag	鞄は少年が持っているか、椅子の上にあるか？
Anaphora （Anaph）	The man holds the bag and the chair; it is green	it の指す対象は鞄か、椅子か？
Ellipsis（Ellip）	The cat chased the mouse, also the rabbit	ウサギは追撃の主体か、対象か？
Adj Scope（Adj）	The woman holds the yellow bag and chair	椅子も鞄と一緒に黄色いか？
Vb Scope（Vb）	And elephant and a bird flying	象は鳥と一緒に飛んでいるか？

像をより細かく分割し、画像中の領域や物体と文中の単語を結びつける手法が提案されている [12, 13]。本研究では、曖昧性が解消された文と画像の特徴量を既存の VLM がどの程度共有しているかを観察し、視覚情報による曖昧性解消能力の現状を探る。

2.2　VLM の弱点：Compositionality

VLM が単に画像の一部を結びつけるだけでなく、文全体の意味を正しく共有できているかを検証するには、文中の述語と項目の連結である Compositionality の観点が重要である。ARO [6] や Winoground [7] は構成単語は同一であるが構文構造が異なる文と画像のペアどうしを区別するタスクを通じて、既存の VLM が文中の単語の繋がりを理解する能力に課題があることを示している。本研究では、これらの評価手法を参考にしつつ、既存の VLM が持つ視覚と言語の曖昧性理解能力に焦点を当てる。特に、曖昧性の理解という、VLM の性能が向上しても依然として課題として残る側面に取り組むことで、モデルの更なる高次の理解能力を引き出すことを目指す。

2.3　視覚情報による曖昧性の解消

視覚情報を用いた曖昧性解消には、実際に曖昧性を含む文と画像のペアを用いる必要があるため、既存の画像・キャプションデータセットの利用が難しい。Berzak らは、曖昧な文と複数の意味を持つ映像・映像からのフレームの画像、構文木などの構造情報を含む LAVA（Language and Visual Ambiguity）コーパスを構築し [1]、後続研究の基盤データとして利用されてきた。しかし、LAVA はデータの量や質に課題があるため、これら問題点を克服する取り組みが報告されている [2, 14, 15]。Mehrabi らは LAVA を補完し、曖昧な文に追加質問を投げて曖昧性を解消する TAB（Text-to-Image Ambiguity Benchmark）

表 2　収集データの統計

	VP	PP	Anaph	Ellip	Adj	Vb	総計
文数	69	50	50	50	50	50	319
ペア数	138	100	100	100	100	100	638

データセットを考案したが [2]、画像データが欠如している点が限界である。一方、Chung らは、画像に基づく翻訳タスクでの曖昧性解消のため、大規模生成モデルを用いて画像と言語データを収集した [3]。本研究では、大規模視覚言語生成モデルを活用し、TAB を基盤に曖昧な文と画像ペアの新たなデータセットを構築することを目指す。

3　データ収集

TAB データセットは曖昧性を 7 種類に分類しているが、一部は言語の曖昧性として適切でなく（付録参照）、また一部はより細かい分類が必要である。本研究では既存の分類を改編し（表 1）、種類ごとに文生成と画像生成の 2 段階でデータ収集を行った。

コーパスの文は TAB データセットを基にしているが、一部に暴力的な単語や実存する人物の名前が含まれており、画像生成モデル [9] によって拒否される可能性がある（付録参照）。本研究は GPT-4 モデル[1)] [8] を用いて文を補完した。収集する文は、曖昧な文とそれを解消した文から構成されており、曖昧な文ごとに二つの意味が含まれる様に設計した（例：The boy holds the green pen and book → The boy holds the green pen and book, the book is not green）。

次に、収集した文を基に DALL-E3 [9] を用いて画像を生成した。生成された画像は作業者の目視による確認を経て選別されたが、そのプロセスにより大量の収集は困難であった。本研究では、簡易的に収集したデータを用いて既存モデルの生成を評価し、今後の課題を示すことに重点を置いた。収集データ

1）　gpt-4o-mini

表 3　曖昧性を解消した文と画像の識別 Accuracy の実験結果

タスク	モデル	VP	PP	Anaph	Ellip	Vb	Adj	全体
T2I	CLIP	0.5	0.46	0.49	0.48	0.53	0.5	0.493
	GPT-4	**0.55**	**0.65**	**0.55**	**0.54**	**0.66**	**0.57**	**0.587**
	LLaVA	0.32	0.32	0.42	0.34	0.41	0.38	0.365
I2T	CLIP	0.5	0.54	0.56	0.59	0.57	0.52	0.525
	GPT-4	**0.88**	**0.95**	**0.79**	**0.62**	**0.93**	**0.9**	**0.845**
	LLaVA	0.45	0.35	0.42	0.42	0.46	0.47	0.428
Dual	CLIP	0.087	0.16	0.18	0.4	0.18	0.12	0.188
	GPT-4	**0.83**	**0.9**	**0.9**	**0.84**	**0.94**	**0.82**	**0.872**
	LLaVA	0.43	0.34	0.36	0.38	0.36	0.48	0.392

表 4　入力（曖昧な元の文と曖昧性を解消した文）による CLIP モデルの両選択肢間の Logit の差（Softmax [16] により%に変換）

	I2T	T2I
曖昧文入力	0.003	0.656
解消文入力	0.241	0.655

の統計は表 2 に示す。

4　実験設定

収集されたデータを用いて、既存の VLM が視覚情報を活用して曖昧性を解消する能力を評価する。

4.1　タスク設定

本研究では、曖昧な文から派生する複数の意味を持つ文群と画像群を正しく対応付ける識別タスクを採用する。このタスクは、似た意味を持つ文と画像を区別する能力を評価するものであり、VLM における細かな意味表現の共有や視覚情報による曖昧性解消能力を測定することが期待される。実験では、識別の正確度（Accuracy）を以下の三つのタスクについて測定する：

- Text-to-Image（T2I）：曖昧性を解消した一文を入力し、二つの候補画像から正しい画像を選ぶ正確度を測定。
- Image-to-Text（I2T）：一枚の画像を入力し、二つの候補文から正しい文を選ぶ正確度を測定。
- Dual：曖昧な文に対応する解消文と画像のペア二組を入力し、両方を正しく対応付けられる正確度を測定。部分的な正解は認めず、完全一致のみを正解とする。

4.2　評価モデル

本実験では、以下のモデルを採用して評価を行う。

- **識別モデル**：識別モデルはプロンプトエンジニアリングを必要とせず、識別確率を直接測定できるため、曖昧性を含む似た意味の理解を評価するのに適している。本実験では、Huggingface の clip-vit-base-patch32 モデル（CLIP [10]）を使用する。
- **生成モデル**：生成 VLM は高い性能を持ち、推論の理由を説明させることでモデルの理解方法を推測することが期待される。本実験では OpenAI の gpt-4o-mini [8] と、LLaVA [17] の 13b モデルを採用する。

5　実験結果・分析

実験結果を表 3 に示す。GPT-4 は全タスクで最も高い性能を示し、LLaVA はプロンプトエンジニアリングをより詳しく設計したのにも拘わらず、タスクの理解が不十分であり、不安定な性能となった。本研究では CLIP と GPT-4 の結果を中心に議論を進める。結果及び分析（図 1）から、曖昧性解消ベンチマークにおける課題の考察を以下の二点に要約できる。

- 視覚情報は言語情報よりも広い埋め込み空間を持つ傾向があり、この差異を克服する方法が求められる。
- モデルの推論及びその結果が識別に繋がる過程には不明瞭な点があり、モデルの理解を徹底的に確認するためのベンチマークが必要である。

5.1　埋め込み空間の差異

表 3 を見ると、CLIP と GPT-4 は共通して T2I より I2T で高い性能を示し、特に GPT-4 では I2T が T2I を大きく上回る。一因として、視覚情報が言語情報より広い埋め込み空間を持ち、画像と言語間の意味関係が一意に決まらないことが挙げられる。表 4 に示す CLIP の Logit 差は、T2I より I2T で大きく、これが主張を裏付けている。図 1 の（C）は GPT-4 の

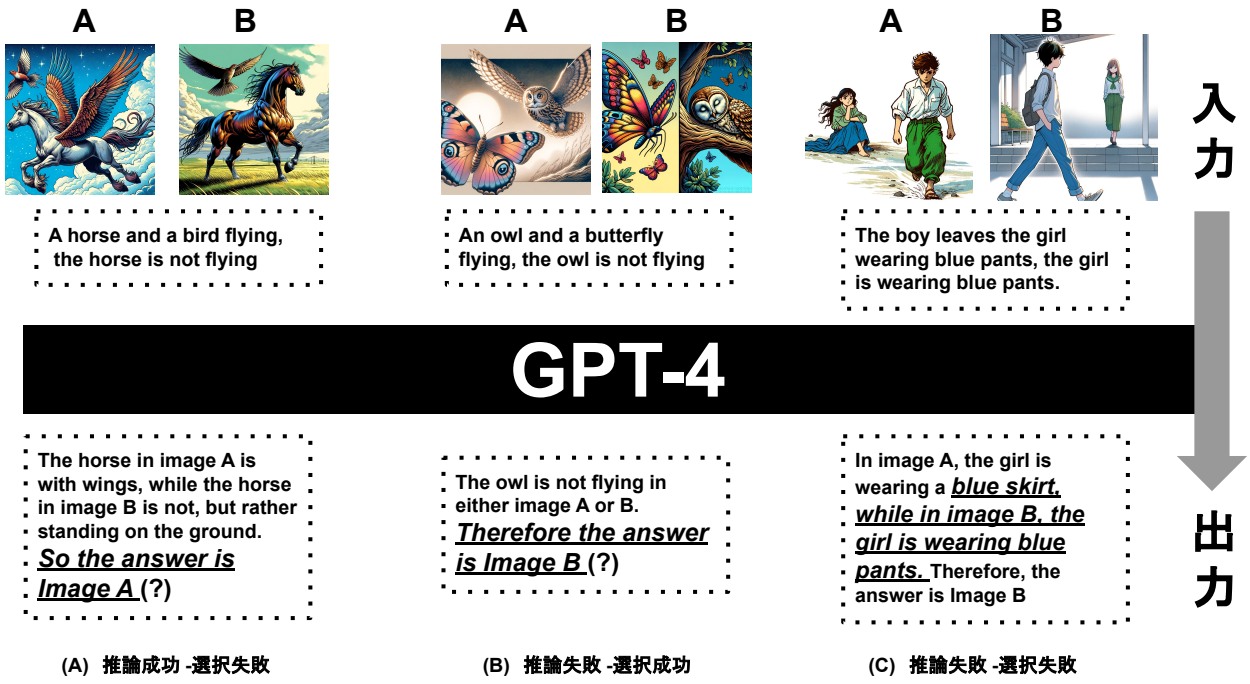

図 1　GPT-4 の実際の推論例

T2I タスクの失敗例で、曖昧性が青いズボンの主体にある。モデルは画像 A のズボンをスカートと誤認し、ズボンの存在だけに集中して誤った画像を選択した。画像のバリエーションがモデルに大きく影響する点は、曖昧性解消タスクにおいて課題となる。これを改善するには、１つの文に対して多様な画像を用意して対応性を高めるか、Scene Graph [18] の様な画像の本質的情報を扱うアプローチが求められる。また、Dual タスクは CLIP の性能が大きく低下したが、これは T2I と I2T を両方考慮する必要があるためと考えられる。一方で、生成モデルの GPT-4 はこのタスクで高い性能を示し、モデル間の原理的違いが反映されている可能性がある。

5.2　モデルの推論過程の透明性

GPT-4 は全体的に高い性能を示しているものの、その推論過程を結果との関連性には疑問が残る。図 1 の（A）は正しい推論をしたにも関わらず誤った選択をした例であり、(B) は誤った推論でありながら正解に至った例を示している。このようなケースは、モデルの推論がどのように識別結果に繋がっているかを十分に把握できていないことを示している。さらに、Dual タスクで GPT-4 が T2I や I2T を上回る結果を示したことは、生成モデルが入力情報の形式や数に大きく影響される可能性を示唆している。この影響を正確に理解しない限り、曖昧性解消の様にモデルの理解工程が鍵となるタスクの実現は困難である。こうした課題に対処するためには、曖昧性解消ベンチマークにおいて、モデルの推論過程を深く分析するための技法が求められる。例えば、VQA（Visual Question Answering） [19] のような質問回答型の分析手法や、Integrated Gradients [20] の様な特徴寄与度の可視化手法を導入することで、モデルの推論と理解の透明性を向上させることが期待される。

6　おわりに

本研究では、視覚情報を用いた曖昧性解消能力を評価及び改善するための新たな枠組みを検討し、簡略に収集したデータを基に既存の VLM を評価した。結果として、GPT-4 が全体的に高い性能を示した一方で、推論課程と結果の関連性に不透明な部分がある可能性を残した。また、視覚情報と言語情報の埋め込み空間の差異や、曖昧性解消タスクにおける生成モデルの影響をより深く理解する必要性も示唆された。今後の研究では、より多様で質量的に優れたデータセットを構築するとともに、モデルの推論課程を詳細に分析することで、曖昧性解消能力の向上を目指す。

謝辞

本研究の一部は JST さきがけ JPMJPR24TC の支援を受けた。

参考文献

[1] Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. Do you see what i mean? visual resolution of linguistic ambiguities. **EMNLP**, 2015.

[2] Ninareh Mehrabi, Palash Goyal, Apurv Verma, J. Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard S. Zemel, A. G. Galstyan, and Rahul Gupta. Resolving ambiguities in text-to-image generative models. **ACL**, 2023.

[3] Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, and Youngjae Yu. Can visual language models resolve textual ambiguity with visual cues? let visual puns tell you! **EMNLP**, 2024.

[4] Ninareh Mehrabi, Palash Goyal, Apurv Verma, J. Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard S. Zemel, A. G. Galstyan, and Rahul Gupta. Is the elephant flying? resolving ambiguities in text-to-image generative models. **ArXiv**, Vol. abs/2211.12503, , 2022.

[5] Ruhiyah Widiaputri, Ayu Purwarianti, Dessi Puji Lestari, Kurniawati Azizah, Dipta Tanaya, and Sakriani Sakti. Speech recognition and meaning interpretation: Towards disambiguation of structurally ambiguous spoken utterances in indonesian. **EMNLP**, 2023.

[6] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Y. Zou. When and why vision-language models behave like bags-of-words, and what to do about it? **ICLR**, 2022.

[7] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. **CVPR**, pp. 5228–5238, 2022.

[8] OpenAI. Gpt-4 technical report. 2023.

[9] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, Aditya Ramesh. Improving image generation with better captions.

[10] 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. **ICML**, 2021.

[11] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. **CVPR**, pp. 10674–10685, 2021.

[12] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel C. F. Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. **CVPR**, pp. 16772–16782, 2021.

[13] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. **CVPR**, pp. 10955–10965, 2021.

[14] Ryosuke Yamaki, Tadahiro Taniguchi, and Daichi Mochihashi. Holographic ccg parsing. **ACL**, 2023.

[15] 李相明, 品川政太朗, 中村哲. Clip におけるテキストの構文情報理解による画像識別能力の向上. 画像の認識・理解シンポジウム. [Lee Sangmyeong, Seitaro Shinagawa, and Satoshi Nakamura (2023). Improving Image Discrimination Ability through Understanding of Textual Syntactic Information in CLIP. Meeting on Image Recognition and Understanding (MIRU)], 7 2023.

[16] John Scott Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. **NeurIPS**, 1989.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **NeurIPS**, 2023.

[18] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. **CVPR**, pp. 1219–1228, 2018.

[19] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. **International Journal of Computer Vision**, Vol. 123, pp. 4 – 31, 2015.

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In **ICML**, 2017.

A　TAB のデータセットからの補完

A.1　曖昧性の種類の再編

TAB [2] のデータセット扱った曖昧性は以下の七種類である：

- Anaphora：代名詞などで前述された対象を指すとき、その候補が複数ある曖昧性（例：the girl looks at the bird and the butterfly; it is red.「赤い」の対象が二つある）
- Ellipsis：省略により文が複数の意味を持つ曖昧性（例：The lion eats the chicken. Also the cat. 猫がチキンを食べるのか、獅子に食べられるのかに分けられる）
- Fairness：キャプションに物体の特性が具体的でないため画像で表現できる候補が複数ある曖昧性（例：the man dusting the floor. 男性の人種、体格などの情報がない）
- Conjunction：接続詞（And、Or 等）で結ばれる複数の名詞に動詞や形容詞が及ぼす影響範囲が複数ある曖昧性（例：the girl hold the green chair and bag「緑」）
- Miscellaneous：物体に複数の性質があるため生じる曖昧性（例：the chicken is ready to eat チキンが食べる行為の主体か対象かに分けられる）
- Syntax-pp：文中で前置詞句が連結される所が複数ある時生じる曖昧性（例：the woman approached the chair with a bag 鞄の位置が女性の腕の中か、それとも椅子の上かに分けられる）
- Syntax-vp：文中で動詞句が連結される所が複数ある時生じる曖昧性（例：the man looked at a boy talking to a telephone 電話をしているのが男性か少年かに分けられる）

この中、Fairness は画像生成の話に関わり、画像の多様性でなく画像が多様性に関わらず含む意味を扱う本研究では適切ではない。それに Miscellaneous は例が稀で、TAB データセットでもサンプル数が三つしかない特殊すぎるケースである。というわけで本研究ではこの二つの種類を除外し、Conjunction を動詞と形容詞の二つに分けて総合 6 種類の曖昧性を扱った。

A.2　TAB データセットの不適切な文の例

TAB のデータセットは、画像生成を想定していながら画像生成モデルの制限に拒まれやすいサンプルが一部あった。まず、kill や threaten、hit などの暴力的な単語があった（例：the girl killed the boy with a gun）。そして現時代の政治家の名前含まれてる例もあり、実際試したところ画像生成モデル [9] に拒絶された（例：Biden sits next to a girl worshiping Trump）。本研究は暴力的な単語を書き換え（例：greet）、実存人物の名前はその特徴を持つ人物の説明に代替した（例：the old man and the blonde man）。