大規模視覚言語モデルに関する指示追従能力の検証

塩野大輝 [1]　宮脇峻平 [1]　田中涼太 [1,2]　鈴木潤 [1,3]
[1] 東北大学　[2]NTT 人間情報研究所　[3] 理化学研究所
{daiki.shiono.s1}@dc.tohoku.ac.jp
{shumpei.miyawaki.b7,tanaka.ryota.r7,jun.suzuki}@tohoku.ac.jp

概要

大規模言語モデル（LLM）の隆盛を背景に，視覚言語モデルに LLM を組み込んだ大規模視覚言語モデル（LVLM）の提案が盛んに行われている．しかし，追加学習後の LVLM は組み込まれる前の LLM が有していた指示追従能力を示さず，タスク指示に従わない事例が観測されている．そこで本研究では，追加学習後の LVLM の指示追従能力が低下することを世界で初めて定量的に示し明らかにする．さらに LVLM の指示追従能力の低下の原因となる要素を洗い出し，モデルの指示追従能力を評価した結果，追加学習時の出力形式に関する指示の有無が指示追従能力の低下に大きな影響を与えている可能性があることを示す．

1　はじめに

視覚情報と言語情報の意味関係を紐づけたマルチモーダルな知識 [1] を活用して，計算機が人により与えられた指示に基づいて適切な推論を実現することは，人工知能研究が目指す最終目的の一つである．2024 年現在では，大規模言語モデル（LLM）の技術的進展が目覚ましく，この進展を受けて，画像とテキストを入力として，テキストを出力する GiT [2]，BLIP-2 [3]，LLaVA [4] などのオープンソースの大規模視覚言語モデル（LVLM）が数多く提案されている．LVLM では言語生成器に LLM を用いることで，多様かつ大規模な言語コーパスで学習された LLM による高度な言語推論能力を活用することができる． LLM / LVLM は，一般に指示文に基づく応答生成ができるほど高いタスク汎化性能を発揮する [5]．また指示文に従う能力は，下流タスクの性能だけでなく，安全性と信頼性という観点からも重要であり，特に医療などのシナリオでは指示に反する意図しない出力が悲惨な結果を招く可能性もある．そのため，指示文に反する出力の軽減とその評価が，安全性・信頼性において重要であると言える．視覚的質問応答タスクなど数多の視覚言語タスクにおいて LVLM は優れた推論能力を実現している．しかし，LVLM に組み込まれる前の LLM であれば指示文に従って回答できるような問題であっても，追加学習後の LVLM は指示文に従うことができない事例が観測されている [6]．

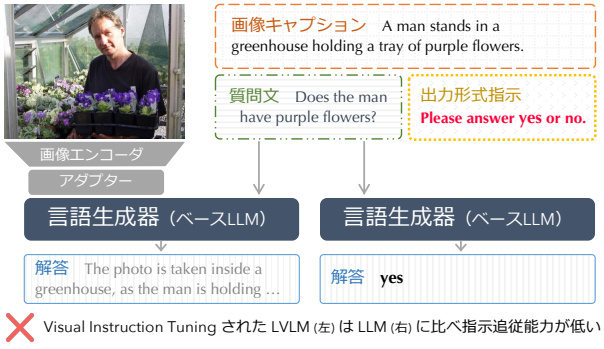

図 1　本研究において取り組む問題の例．LVLM の指示追従能力が LLM に比べて低いという現状を定量的に明らかにし，指示追従能力の低下をもたらす要因を考察する．

本研究では，追加学習後の LVLM の指示文に従う能力を定量的に評価し，指示文に従う能力が低下していることを初めて定量的に示し明らかにする．さらに，LVLM の指示文に従う能力低下に対する原因の解明を目指す（図 1）．まず，出力形式に関する指示を含んだ追加学習データセットを新たに作成する（3 章）．さらに，追加学習後の LVLM の指示追従能力の低下を定量的に示す（4, 5 章）．加えて，LVLM の指示追従能力の低下の原因となる要素を洗い出し，モデルの指示追従能力を評価することで，特定の要素が LVLM の指示追従能力に与える影響を調査する．実験結果より，LVLM の指示追従能力の低下に大きな影響を与えているのは，追加学習時の出力形式に関する指示の有無である可能性が高く，視覚情報による影響は小さい可能性があることを示す．

2 関連研究

2.1 LLM / LVLM における指示追従能力

人間から与えられた指示に従った応答文を LLM が生成することは，意図しない応答文の生成を抑制するという観点から社会的実装において重要な役割を持つ．Li らは LLM の “指示追従能力” を，学習データセットから取得可能な文脈知識とは異なる指示文に従う能力とみなし，指示追従能力の評価を試みている [7]．Li らに倣い，本研究では，指示追従能力を「モデル（LLM / LVLM）が入力された事前知識と一致しない可能性のある指示文に従う能力」と定義し，モデルの指示追従能力の評価を実施する．近年では複雑な人間の指示に高い追従性を示す LLM が数多く提案されており，これらの LLM は多様なタスクに対して，指示文と対象タスクの入出力ペアが対になった指示付きの学習データセットで追加学習（Instruction Tuning（指示学習））される．また指示学習の入力データに画像を含めることで，指示追従能力の獲得が視覚言語タスクまで拡張された Visual Instruction Tuning（視覚指示学習）も提案されており [4]，LVLM の学習に広く使用されている．

2.2 LVLM の指示追従能力の脆弱性

優れた汎化性能を実現する LVLM であるが，画像中に存在する物体の位置や数を正確に把握できないこと，架空の物体に関する指示に対して誤認した応答を生成することなど，視覚処理および指示追従の観点の一部で脆弱性が定性的に観察されている [6]．しかし，LVLM の指示追従能力の定量的な評価は依然としてなされていない．そこで本研究では，LVLM の指示追従能力を定量的に評価する．

2.3 視覚指示学習データセットの課題

先行研究にみられる視覚指示学習データセット [4, 8] では，GPT-4 [9] のような LLM を用いて画像キャプションから指示付きの学習データを生成する．しかし，我々はこれらの視覚指示学習データセットには，出力形式に関するタスク指示が含まれてないことを定性的に確認した．そこで本研究では，出力形式に関するタスク指示が含まれた（視覚）指示学習データセットを新たに作成し，出力形式に関するタスク指示の有無による LVLM の指示追従能力に対する影響を調査する．

3 指示追従データセット

3.1 （視覚）指示学習データセットの作成

視覚指示学習データの出力形式に関する指示の不足による影響 図 1 では，LVLM が出力形式に関する指示内容を正しく考慮できていない例を示した．を対象とした研究分野において，多様性を担保しながら出力形式を明示的に指示するようなデータセットは十分に存在しない，そこで我々は追加学習用の二種類の視覚指示学習データセットを作成した（図 2）．まず COCO 2014 [10] の検証データセットから 10,000 件の画像を無作為に収集した．次に，これら 10,000 件の画像全てに対して，GPT-4V [11] を使用して画像キャプションを生成した．また，LLM の指示追従能力を評価するためのデータセットである IFEval データセット [12] から出力形式に関するタスク指示を 100 件無作為に抽出した．さらに画像キャプションと出力形式に関するタスク指示から質問文と解答を GPT-4 により生成した．これにより，{ **画像**, **画像キャプション**, **出力形式に関するタスク指示**, **質問文**, **解答** } のペアを 5,000 件得た．これにより，以下の二種類のデータセットが作成できる．

- Format Oriented Visual Instruction Tuning (FOVIT) データセット（図 2.a）：{ **画像**, **出力形式に関するタスク指示**, **質問文**, **解答** } のペア 5,000 件
- Format Oriented Instruction Tuning (FOIT) データセット（図 2.b）：{ **画像キャプション**, **出力形式に関するタスク指示**, **質問文**, **解答** } のペア 5,000 件

視覚情報による影響 追加学習時に入力する情報を言語情報のみに限定した場合，すなわち LVLM のベース LLM をテキストのみで追加学習する場合，出力形式に関する指示の有無がベース LLM の指示追従能力に与える影響を調査する．具体的には，出力形式に関する指示が含まれたデータセットで追加学習した LLM と出力形式に関する指示が含まれないデータセットで追加学習した LLM を 3.2 節で作成した指示追従能力評価データセットで評価する．3.1 節の結果と本節の結果から，視覚情報の有無がモデルの指示追従能力に与える影響を調査できる．上述の画像キャプションと質問文から解答を GPT-4 により生成し，{ **画像**, **画像キャプション**, **質問文**, **解答** } のペアをさらに 5,000 件得た．これにより，以

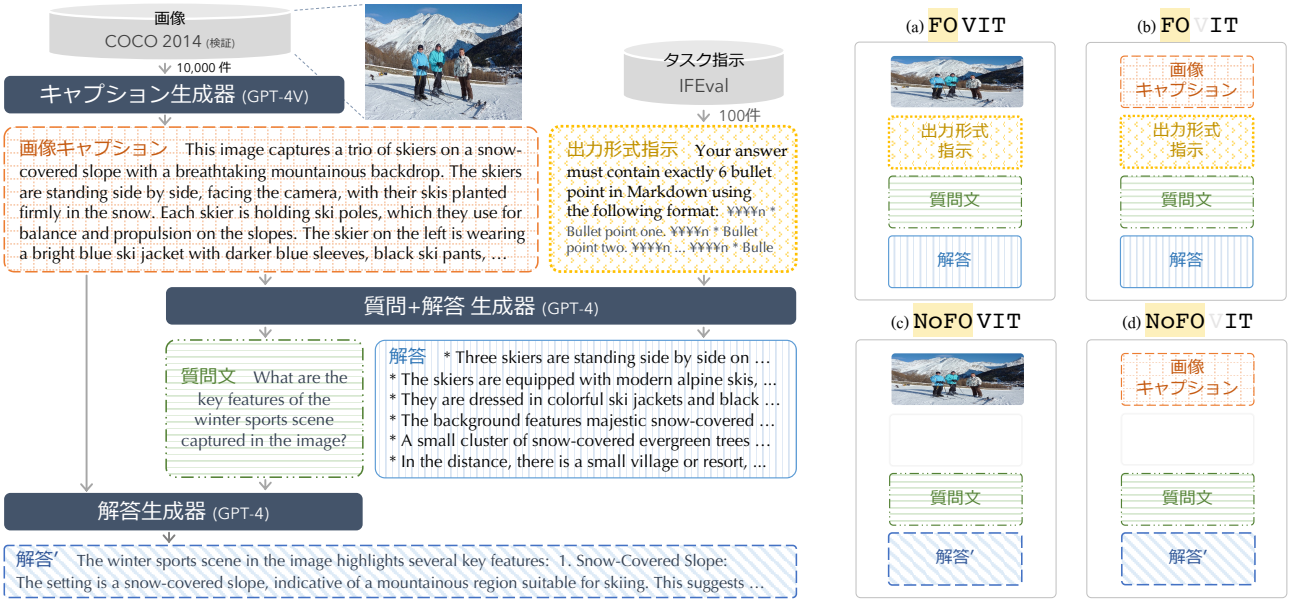

図 2　COCO 画像と，GPT-4V により生成された画像キャプション，IFEval データセットから抽出された出力形式に関するタスク指示，GPT-4 により生成された質問文と解答からなる（視覚）指示学習データセットの一例（3.1 節）．

下二種類のデータセットが作成できる（図 2）．

- Not Format Oriented Visual Instruction Tuning (NoFOVIT) データセット（図 2.c）：{ **画像**, **質問文**, **解答** } のペア 5,000 件
- Not Format Oriented Instruction Tuning (NoFOIT) データセット（図 2.d）：{ **画像キャプション**, **質問文**, **解答** } のペア 5,000 件

3.2　指示追従能力評価用データ

Li ら [7] は，LLM の指示追従能力を評価するにあたり，事前に定義されたラベルの生成が必要な二値分類タスクを評価対象として，LLM が指示文に従うかどうかを評価する verbalizer manipulation を提案している．例えば感情分類では positive，negative というラベルが定義されるが，**分類タスクを解くという指示に従う**のであれば a，b のようなラベルを割り当てることで十分である．Verbalizer manipulation では，ラベルの意味表現と学習時の文脈知識の整合性別に以下の三つのラベル体系を定義して，LLM が分類タスクを解くという指示に追従するか段階的に評価する（図 3）：

- **Natural**: ラベルと学習時の文脈の意味表現が一致する（整合性が高い）．
- **Neutral**: ラベルと学習時の文脈の意味表現に関連性がない．
- **Unnatural**: ラベルと学習時の文脈の意味表現が一致しない（整合性が低い）．

評価データセットの事例

If a movie review is **positive**,
you need to output "{label_0}".
If a movie review is **negative**,
you need to output "{label_1}".

Movie review: lovely and poignant.

Answer:

文脈との整合性別 ラベル体系		If positive.. label_0	If negative.. label_1
Natural	高	positive	negative
Neutral	↕	foo	bar
Unnatural	低	negative	positive

図 3　SST-2 のデータを用いた verbalizer manipulation による評価セットの作成方法．文脈とラベルの意味表現との整合性に基づいて “Natural”，“Neutral”，“Unnatural” の 3 種類のラベルが定義される．

Verbalizer manipulation によって，モデルが事前知識に依存しているか，または指示に正確に従うために事前知識を上書きする能力（指示追従能力）を評価することができる．

我々は，Li ら [7] に倣い，九つの二値分類データセット（SST-2 [13]，FP [14]，EMOTION [15]，SNLI [16]，SICK [17]，RTE [18]，QQP [19]，MRPC [20]，SUBJ [21]）それぞれに対して 12 セットの verbalizer manipulation を実施して評価データセットを構築した．

4 実験設定

視覚指示学習が LVLM の指示追従能力にもたらす影響を調査すべく，LVLM およびその言語生成器であるベース LLM を 3.1 節のデータセットで学習し，各モデルの指示追従能力を 3.2 節のデータセットを用いて評価を行った．

評価対象となるモデル LVLM，LLM の追加学習前のベース LLM には Llama 2-Chat 7B [22] を使用した．追加学習済みのモデルは $\text{LVLM}_{\text{FOVIT}}$，$\text{LVLM}_{\text{NoFOVIT}}$，$\text{LLM}_{\text{FOIT}}$，$\text{LLM}_{\text{NoFOIT}}$，$\text{LVLM}_{\text{LLaVA}}$ で表現し，それぞれ `FOVIT`, `NoFOVIT`, `FOIT`, `NoFOIT`, `LLaVA-Instruct-150K` で追加学習したことを示す．LVLM，LLM の追加学習時の詳細な設定は付録を参照されたい．推論時は全てのモデルで最大出力トークン数を 15，ビーム幅を 10 とした．

評価データセット 3.2 節で作成した指示追従能力評価データセットで評価した．このデータセットは，九種類の二値分類タスクを対象に図 3 のような verbalizer manipulation を実施したものである．詳細なデータ件数は付録に示した．LVLM を 3.2 節で作成したデータセットで評価する際には，該当する画像情報が存在しないため，白塗りの画像を視覚情報として入力した．

評価指標 評価指標には，`SQuAD v2` [23] で提案された F_1 値を採用した．本指標では，事前に定義した二値分類のラベルを正解のトークン系列とする．言語モデルによって生成された応答トークン系列と正解のトークン系列との重複トークン数に基づいて再現率および適合率が算出され，正解と重複するトークン数が多いほど F_1 値が大きくなる．

5 実験結果

3.2 節によって作成されたデータセットを用いて，追加学習済みのモデルの指示追従能力を評価した結果を図 4 に示す．

指示追従能力の低下の検証 Verbalizer manipulation における Natural, Neutral および F_1 値のマクロ平均を示す All において，追加学習を実施した $\text{LVLM}_{\text{FOVIT}}$ を除く全てのモデルがベース LLM である LLM (Llama 2-Chat 7B) よりも低い F_1 値となった．特にモデルが事前知識に依存せずに，与えられた指示に従うかを評価する Unnatural においては全ての追加学習済みモデルにおいて LLM (Llama 2-Chat 7B) を下回る結果となった．このことは，ベース LLM が有していた指示追従能力の低下が追加学習に起因することを示唆している．

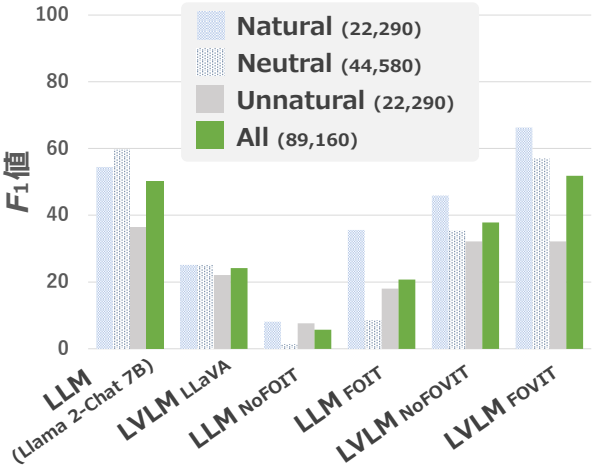

図 4 評価データセットに対する F_1 値．$\text{LVLM}_{\text{FOVIT}}$，$\text{LVLM}_{\text{NoFOVIT}}$，$\text{LLM}_{\text{FOIT}}$，$\text{LLM}_{\text{NoFOIT}}$，$\text{LVLM}_{\text{LLaVA}}$ はそれぞれ `FOVIT`, `NoFOVIT`, `FOIT`, `NoFOIT`, `LLaVA-Instruct-150K` データセットで追加学習したモデルを表す．また “All” は “Natural”，“Neutral”，“Unnatural” の F_1 値のマクロ平均．

視覚指示学習データ中の出力形式に関する指示の有無による影響 Verbalizer manipulation における全てのラベル体系において，$\text{LVLM}_{\text{NoFOVIT}}$ よりも $\text{LVLM}_{\text{FOVIT}}$ の F_1 値が高い結果となった．FOVIT は出力形式の指示が含まれた視覚指示学習データセットであることから，出力形式を明示的に与えることによりベース LLM が有している指示追従能力の低下を抑制できることが示唆された．

視覚情報による影響 また全てのラベル体系において，$\text{LLM}_{\text{NoFOIT}}$ よりも LLM_{FOIT} の F_1 値が高い結果となった．FOIT は出力形式の指示が含まれた指示学習データセットであることから，追加学習時に入力する情報を言語情報のみに限定した場合であっても，出力形式を学習データ中に明示的に与えることによってベース LLM が有している指示追従能力の低下を抑制できることが示唆された．

6 おわりに

本研究では，LVLM の指示追従能力を定量的に評価した．実験結果より，LVLM の指示追従能力の低下をはじめて定量的に確認できた．さらに，LVLM のみならず LLM においてもモデル追加学習時の出力形式に関する指示の有無が指示追従能力に大きな影響を与える可能性があることを示唆した．今後の展望として，出力形式に関する指示データの量とモデルの指示追従能力との関係を精緻に調査したい．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research) の助成を受けて実施されたものである．研究遂行にあたりご助言ご協力を賜りました Tohoku NLP グループの皆様に感謝申し上げます．

参考文献

[1] C.K. Odgen and I.A. Richards. **The Meaning of Meaning A Study of the Influence of Language upon Thought and of the Science of Symbolism**. Routledge & Kegan Paul Ltd., 1923.

[2] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. **arXiv preprint arXiv:2205.14100**, 2022.

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. **arXiv preprint arXiv:2301.12597**, 2023.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **arXiv preprint arXiv:2304.08485**, 2023.

[5] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In **International Conference on Learning Representations**, 2021.

[6] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. **arXiv preprint arXiv:2306.13394**, 2023.

[7] Shiyang Li, Jun Yan, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, and Hongxia Jin. Instruction-following evaluation through verbalizer manipulation. In **NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following**, 2023.

[8] Bo Zhao, Boya Wu, and Tiejun Huang. SVIT: scaling up visual instruction tuning. **CoRR**, 2023.

[9] OpenAI. Gpt-4 technical report, 2023.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In **Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13**, pp. 740–755. Springer, 2014.

[11] OpenAI. Gpt-4v(ision) system card, 2023.

[12] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. **arXiv preprint arXiv:2311.07911**, 2023.

[13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In **Proceedings of the 2013 conference on empirical methods in natural language processing**, pp. 1631–1642, 2013.

[14] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. **Journal of the Association for Information Science and Technology**, Vol. 65, No. 4, pp. 782–796, 2014.

[15] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In **Proceedings of the 2018 conference on empirical methods in natural language processing**, pp. 3687–3697, 2018.

[16] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. **arXiv preprint arXiv:1508.05326**, 2015.

[17] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014**, pp. 216–223. European Language Resources Association (ELRA), 2014.

[18] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In **Machine learning challenges workshop**, pp. 177–190. Springer, 2005.

[19] Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. Quora question pairs, 2017.

[20] Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In **Third International Workshop on Paraphrasing (IWP2005)**, 2005.

[21] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. **arXiv preprint arXiv:1803.05449**, 2018.

[22] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[23] Pranav Rajpurkar, Jian Zhang, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In **ACL 2018**, 2018.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pp. 8748–8763. PMLR, 2021.

A 参考情報

A.1 指示追従能力評価データセットの種類別データ件数

表 1 評価データセットの種類別データ件数

データセット名	ラベルタイプ		
	Natural	Neutral	Unnatural
SST-2 [13]	2,616	5,232	2,616
FP [14]	2,619	5,238	2,619
EMOTION [15]	3,000	6,000	3,000
SNLI [16]	3,000	6,000	3,000
SICK [17]	3,000	6,000	3,000
RTE [18]	831	1,662	831
QQP [19]	3,000	6,000	3,000
MRPC [20]	1,224	2,448	1,224
SUBJ [21]	3,000	6,000	3,000
合計	22,290	44,580	22,290

3.2 節で構築した指示追従能力評価データセットのデータセット名とラベルタイプの種類別データ件数を表 1 に示す．

A.2 追加学習時の学習設定

表 2 LVLM の追加学習時の学習設定

エポック数	1
グローバルバッチサイズ	32
最適化関数	AdamW
初期学習率	5.0×10^{-5}
スケジューラー	cosine
最大系列長	512
画像エンコーダ	CLIP ViT-Large/14 [24]
アダプター	1 層の線形層
LLM	Llama 2-Chat 7B [22]
学習可能なパラメータ	{アダプター, LLM}

表 3 LLM の追加学習時の学習設定

エポック数	1
グローバルバッチサイズ	32
最適化関数	AdamW
初期学習率	5.0×10^{-5}
スケジューラー	cosine
最大系列長	1024
LLM	Llama 2-Chat 7B

4 章で説明した LVLM，LLM の追加学習時の設定をそれぞれ表 2，3 に示す．

A.3 評価データセットにおけるモデル出力の定性評価

5 章において，評価データセット（SST-2 の “Natural” タイプ）の一例に対する複数モデルの出力結果を比較したものを図 5 に示す．出力形式の指示

```
You are a helpful assistant judging the sentiment of a movie
review. If the movie review is positive, you need to output
"positive". If the movie review is negative, you need to
output "negative". You are only allowed to output "positive"
or "negative".
Movie review: there 's ... tremendous energy from the cast ,
a sense of playfulness and excitement that seems appropriate.
Answer (Gold): positive

LVLM FOVIT: positive
LVLM NoFOVIT: The sentiment of the movie review is positive
LLM FOIT: positive
LLM NoFOIT: Yes, the movie review is positive
LVLM LLaVA: The sentiment of the movie review is positive
LLM (Llama 2-Chat 7B): positive
```

図 5 評価データセット（SST-2 の “Natural” タイプ）に対する複数モデルの出力結果の一例．

付きのデータセットで LLM (Llama 2-Chat 7B) を追加学習したモデル $\text{LVLM}_{\text{FOVIT}}$，$\text{LLM}_{\text{FOIT}}$ は与えられた指示に従い正解することができている．一方で，出力形式の指示が含まれないデータセットで追加学習したモデル $\text{LVLM}_{\text{NoFOVIT}}$，$\text{LLM}_{\text{NoFOIT}}$，$\text{LVLM}_{\text{LLaVA}}$ は，指示に従うことができていない．この結果からも，出力形式の指示が含まれたデータセットの利用が，追加学習前のベース LLM の指示追従能力低下の抑制に役立つ可能性が示唆された．

A.4 本研究の限界

3.1 節で作成された追加学習用のデータセットは，GPT-4V により画像キャプションを生成していたため，画像の内容と一致しない情報が含まれてしまう可能性がある．また 5 章では，モデルの指示追従能力をラベルとの一致という限定的な方法のみで評価しているので，包括的に指示追従能力を測れているとは言えない．加えて LVLM の評価の際は，理想的には，視覚情報を参照しながら指示追従能力を評価したいが，現在の方法では，視覚情報を白塗りの画像で補完してしまっており，テキスト情報だけに手がかりがある評価方法になってしまっている．