大規模視覚言語モデルの質感知覚能力の分析

松田 陵佑 [1]　塩野 大輝 [1]　Ana Brassard[2,1]　鈴木 潤 [1,2,3]
[1] 東北大学　[2] 理化学研究所　[3] 国立情報学研究所
{matsuda.ryosuke.t4,daiki.shiono.s1}@dc.tohoku.ac.jp
ana.brassard@riken.jp　jun.suzuki@tohoku.ac.jp

概要

本研究では，「質感」に焦点を当て，大規模視覚言語モデル（LVLM）の質感知覚能力を調査し，さらに LVLM と人間との間の質感知覚の整合性を分析することを目的とする．はじめに画像内の物体に対して人間が知覚する質感語を人手で収集した．次に，収集した質感語をもとに，LVLM が適切な質感語を選択できるか評価する分類タスクを設計し，LVLM と人間の正解率を算出した．また，LVLM に質感語を生成させ，その出力を人間が評価する生成タスクも実施した．最終的には，分類タスクの正解率が高い LVLM は，生成タスクにおいても高いスコアを示すことを確認し，分類タスクが，LVLM の質感知覚能力の評価だけでなく，人間知覚の整合性まで簡易に評価できる可能性があることを示す．

1　はじめに

人間が持つ感覚に関して，機械学習により獲得したモデルと人間との整合性について，多元質感知 [1] に注目した整合性，多感覚整合性 [2] といった研究で分析や議論がおこなわれてきた． モデルと人間との間の整合性が取れていることは，モデルが人間と同様の方法で知覚し，思考し，行動する際に重要となる [3, 4, 5]．本研究では，「質感」に焦点を当て，大規模視覚言語モデル (LVLM) と人間との間の質感知覚能力の整合性を調査することを目的とする．質感は，人によって異なる解釈が存在するため，明確に定義することが難しい．さらに質感という概念 [6] が何を包含しているのか，その境界は曖昧かつ主観的である．本研究では，質感を下記 3 つの概念を包含するものとして定義している [7, 8].

- 物性（光沢感・透明感など）
- 状態（乾燥・凍結など）
- 印象（美しい・醜いなど）

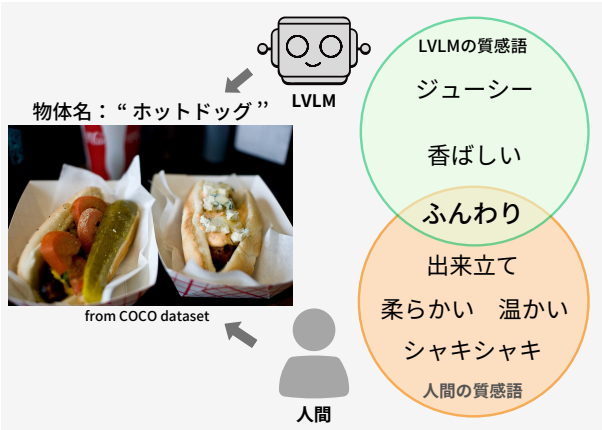

図 1　LVLM が出力する質感語と人手で収集した質感データをもとに LVLM の質感知覚と人間の質感知覚の整合性を評価する．

本研究では，画像中の物体に対して人間が感じ取るような質感を LVLM がどの程度知覚しているかを分析するために，{ 画像, 物体名, 質感語 } の 3 つの情報を結びつけた質感データセットを人手で作成（2 章）し，そこから分類タスクと生成タスクの 2 種類のタスクを設計する．分類タスクは，提示された画像と画像内の物体に対して最も適切な質感語を選択する問題となっており，このタスクを用いて，既存の代表的な LVLM が有する質感知覚能力を評価する（3 章）．また生成タスクでは，COCO データセット [9] の画像に含まれている物体に対して，LVLM に質感語を生成させる．そして，LVLM が生成した質感語が人間から見て自然であるかを人手評価することで，LVLM の質感知覚と人間の質感知覚との整合性を調査する．これに加えて，LVLM が生成した質感語と質感データセット中の人間が書き出した質感語の共通部分を質感語一致率と定義し，評価する（4 章）．最終的に，LVLM の分類タスクと生成タスクの実験結果から，分類タスクが LVLM の質感知覚能力の評価だけでなく，人間知覚の整合性まで簡易に評価できる可能性があることを示す（5 章）．

2　質感データセットの作成

本研究の分析対象となる質感語を含むデータセットは存在していなかったので，我々は，質感データセットを新たに構築した．まずベースとなる画像データとして，COCO データセット[1)]を選択した．COCO データセットからサンプリングした画像内の物体に対して，クラウドワーカーによる質感語の人手アノテーション[2)]により，{ 画像, 物体名, 質感語 } の 3 つ組を 1 サンプルとした合計 22,409 件のサンプルからなるデータセットを作成した．質感データセット構築の際，データの質を向上させるために，画像全体に対する画像内の物体が占める面積の割合が 15% 以下となるサンプルは除去した．

3　分類タスク

LVLM の質感知覚能力を分析するために，2 章で説明した質感データセットを基にして，適切な質感語を選択する分類タスクを設計し，評価する．

3.1　分類タスクの設計

質感データセットから，画像と物体名が同一で質感語のみが異なる 3 つのサンプルを抽出し，これ以外のサンプルから 2 つの質感語を無作為に抽出する．これにより，画像と物体名に対して，自然な可能性の高い質感語 3 つと無関係な可能性の高い質感語を 2 つ抽出する．そして，1 つの画像と物体名ペアに対して，5 つの質感語をそれぞれ 5 名のアノテーターに提示し，画像と物体名ペアに対して，その質感語が適切か不適切か 2 択で回答してもらう．5 名のアノテーターの内，4 名以上が適切であると回答した画像と物体名をその質感語の正例とみなす．このプロセスにより，画像と画像内の物体に対する質感語をより妥当性の高いものにしている．最終的に，N 択問題に対して，質感語の正例 1 つと，負例 $N-1$ つ抽出することで，分類タスクを設計した．

3.2　実験設定

分類タスクでは，画像内の物体に対して与えられた複数の質感語候補の中から最も適切なものを 1 つ選択させることで正解率 (%) を算出し，LVLM が質感を正しく理解しているかどうかを明らかにする．

1)　COCO 2017 の訓練 (Train)/検証 (Val) 用画像データを使用．
2)　人手アノテーションには，Yahoo!クラウドソーシングを用いており，1 章で説明した質感語の定義をワーカーに明示した上で，各画像と画像内の物体に対する質感語を収集した．

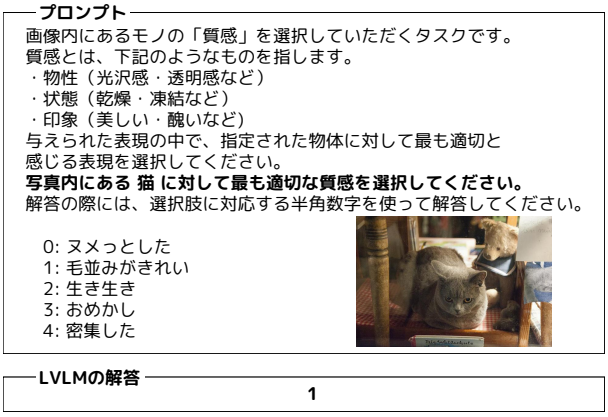

図 2　LVLM に与える分類タスクのプロンプトとその入出力例．入力画像の例は右下の「猫」の画像である．

表 1　各 LVLM と人間の分類タスクの正解率 (%) の結果．表中のモデル名からモデルが一意に定まらないものに関しては，脚注[3)]に使用したモデルの配布元 URL を記した．また，Human のスコアは，3 名のアノテーターの平均正解率である．

モデル	2 択問題	5 択問題
Random	50.00	20.00
Human	–	78.57
GPT-4o$_{\text{2024-11-20}}$ [10]	93.43	81.19
Llama-3.2 11B$_{\text{Instruct}}$ [11]	57.31	45.07
Qwen2-VL 7B$_{\text{Instruct}}$ [12]	85.37	61.79
LLaVA-OneVision 7B$_{\text{OV}}$ [13]	78.21	62.09
LLaVA-NeXT 7B [14]	64.48	33.43
Idefics2 8B$_{\text{chatty}}$ [15]	66.27	39.40
LLaVA-1.5 7B [16]	52.84	21.49

本実験では，3.1 節で説明した設計方法に基づいて，選択肢が 2 択と 5 択の分類タスクをそれぞれ 335 件用意した．実際に与えるプロンプトとその回答形式は，図 2 に示した．これら 2 択と 5 択の分類タスクをそれぞれ表 1 に示した 7 種類の LVLM と 3 名のアノテーターに回答してもらい，評価を実施する．

3.3　実験結果

分類タスクを 7 種類の LVLM と 3 名のアノテーターに解かせて算出した正解率を表 1 [3)]に示した．2 択，5 択問題の両方で，最も正解率の高い LVLM は GPT-4o [10] であり，人間のスコアを 2.62%上回る結果となった．また，GPT-4o は 2 択問題よりも

3)　表 1 から，モデル名が一意に定まらないモデルの配布元 URL を以下に示す．
LLaVA-OneVision-7B$_{\text{OV}}$:https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-chat-hf,
LLaVA-NeXT-7B:https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

プロンプト

画像内にあるモノの「質感」を述べていただくタスクです。
質感とは、下記のようなものを指します。
・物性（光沢感・透明感など）
・状態（乾燥・凍結など）
・印象（美しい・醜いなど）
画像に対して感じる質感を表す表現を 3 個「、」で分けて書いてください。
その際、3個とも出来るだけ違う観点の表現にしてください。
また、似た表現（きらきら・きらんきらんなど）にならないようにしてください。

写真内にある「果物」は、どのような質感を持っているように感じますか？

LVLMの回答

みずみずしい、カラフル、甘い

図 3　LVLM に与える生成タスクのプロンプトとその入出力例．入力画像の例は右の「果物」の画像である．

選択肢が多く難しいタスク設定の 5 択問題に関しても Rondom スコアを大きく上回る正解率を示し，選択肢の数に対して頑健であることが分かる．また，比較的新しいオープンソースモデルである，Qwen2-VL 7B$_{\text{Instruct}}$ [12] や LLaVA-OneVision 7B$_{\text{OV}}$ [13] も人間のスコアに近づきつつある．一方で，比較的古い LVLM である LLaVA-1.5 7B [16] は，2 択問題と 5 択問題の両方のタスクにおいて，チャンスレートとほぼ同程度の正解率しか示さなかった．

4　生成タスク

LVLM と人間の質感知覚の整合性を分析するために，画像内の物体に対して，LVLM が生成した質感語と人間が書き出した質感語を用いて，与えられた質感語が人間にとって自然であるかを評価する．

4.1　生成タスクの設計

2 章でアノテーターに画像と物体名から質感語を書き出してもらう際に提示した説明文とほとんど同様の説明文をプロンプトとして LVLM に与え，質感語を生成させる．図 3 に示すように，LVLM に与えるプロンプトには，質感語生成に関するタスク指示と画像と特定の物体名の情報を含む．そして，全 107 件のサンプルを対象に，画像と物体名のペア 1 つにつき，LVLM に質感語を 3 つずつ生成させる．

4.2　実験設定

画像と物体名から，LVLM が生成した質感語と 2 章で人間が書き出した質感語との間の整合性を調査するために，LVLM が生成した質感語が自然かどうか人手評価する yes/no 人手判定と LVLM が生成した質感語と人間が書き出した質感語との質感語一致率の 2 つの指標による評価を実施する．

図 4　LVLM が生成した質感語と 2 章で人間が書き出した質感語の積の算出の様子．入力画像の例は右の「ホットドッグ」の画像である．図例では，LVLM の質感語の出力が 3 つあり，そのうち「ふんわり」が人間の回答と一致しており，質感語一致率の計算をする際の $y_{h,i}$ は 1，w_i は 3 ということになる．

yes/no 人手判定　画像と物体名に対して，LVLM が生成した 3 つの質感語それぞれに対して，3 名のアノテーターが自然な質感表現であるか否かを yes/no で判定する．

$$\text{yes/no 人手判定スコア} = \frac{1}{H}\sum_{h=1}^{H}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{y_{h,i}}{|w_i|}\right) \quad (1)$$

ただし，H を全アノテーター数（3 名），N を全評価サンプル件数（107 件），$y_{h,i}, w_i$ をそれぞれ評価サンプル i においてアノテーター h が yes と回答した数及び LVLM が生成した質感語の数（3 つ）を表す．

質感語一致率　画像と物体名に基づき，LVLM が生成した 3 つの質感語と 2 章で人間が書き出した質感語の共通部分を求める（図 4）．そして，各サンプルごとに，LVLM が生成した質感語集合 S_{LVLM} と人間が書き出した質感語集合 S_{Human} との間の質感語一致率を算出し，平均する．

$$\text{平均質感語一致率} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|S_{\text{Human}} \cap S_{\text{LVLM}}|}{|S_{\text{LVLM}}|}\right) \quad (2)$$

ただし，LVLM の質感語生成時には句読点区切りで 3 つの質感語を生成するようにプロンプトを設計していたが，この条件を満たしていない出力は質感語一致率の計算から除外する．

アノテーション作業のコストを鑑みて，表 1 に記した 5 択分類問題の評価結果に関して，最も正解率の高かった GPT-4o と最も低かった LLaVA-1.5 7B の 2 種類に対して，生成タスクの評価を実施した．

4.3　実験結果

評価対象の LVLM として，GPT-4o 及び LLaVA-1.5 7B を選択し，全 107 件のサンプルを対象に，画像と物体名のペア 1 つにつき，質感語を 3 つずつ生成さ

表 2　全 107 件のサンプルに対して，LVLM が生成した 3 つの質感語が自然か否かをそれぞれ 3 名のアノテーターが yes/no 人手判定した結果.

モデル	yes/no 人手判定スコア
GPT-4o	75.49
LLaVA-1.5 7B	57.75

表 3　全 107 件のサンプルに対して，LVLM が生成した質感語集合と人間が書き出した質感語集合との間の質感語一致率を算出し，平均した結果.

モデル	平均質感語一致率
GPT-4o	21.50
LLaVA-1.5 7B	11.93

せた．さらに，生成された質感語が人間にとって自然であるかを「yes/no 人手判定」と「質感語一致率」の 2 つの指標で評価した．

yes/no 人手判定　GPT-4o 及び LLaVA-1.5 7B が生成した質感語に対して，3 名のアノテーターにより，yes/no 判定を実施し，式 (1) に基づいて，yes/no 人手判定スコアを算出した結果を表 2 に示した．表 2 から，LLaVA-1.5 7B よりも GPT-4o が出力した質感語の方が，17.74% 程度人間が自然であると判断したことが分かる．この結果から，分類タスクにおける正解率が高い LVLM は，yes/no 人手判定においても高い評価値となる可能性が示唆された．また，GPT-4o 及び LLaVA-1.5 7B が生成した質感語の内，3 名のアノテーター全員が no を選択したサンプルは，全 107 件のサンプルの内 6 件あった．この 6 件のサンプルの内，4 件のサンプルを図 5 に示した．図 5 に示したサンプルから，LVLM は，画像情報を参照せずに，物体名だけから質感語を生成してしまう場合がある可能性が示唆された．

質感語一致率　句読点区切りの生成に失敗している質感語に関しては，質感語一致率の計算から除いたため，結果として，GPT-4o は 106 件，LLaVA-1.5 7B は 81 件のサンプルに対して式 (2) に基づいて，平均質感語一致率を算出した．その結果を表 3 に示した．表 3 から，LLaVA-1.5 7B より GPT-4o が生成した質感語の方が，9.57% だけ人間が書き出した質感語と一致している割合が多いことが分かる．この結果から，分類タスクにおける正解率が高い LVLM は，質感語一致率においても高い評価値となる可能性が示唆された．

図 5　GPT-4o 及び LLaVA-1.5 7B が生成した質感語の内，3 名のアノテーター全員が不自然（no）と選択したサンプル { 画像, 物体, 質感語 } の例.

5　分析と議論

生成タスクの yes/no 人手判定では，人間が直接判定を行うので，LVLM と人間の質感知覚の整合性を厳密に評価できる利点があるが，評価に人間が介入するため，評価時の人的コストが高く，評価の実施難易度が高い．一方で，分類タスクは，複数の選択肢の中から画像内の物体に適した質感語を 1 つ選択するタスクであり，人手を介さずに正解率を容易に算出できる．また，4.3 節の実験結果から，分類タスクにおける正解率が高い LVLM は，生成タスク（yes/no 人手判定，質感語一致率）のスコアも高くなっており，分類タスクと生成タスクのスコアは相関がある可能性が示唆される結果が得られていた．このことから，本来，生成タスクを用いなければ明らかにならない，LVLM と人間との間の質感知覚の整合性の評価を，我々が設計した分類タスクを用いることで間接的に評価できる可能性がある．

6　おわりに

本研究では，代表的な LVLM の質感知覚能力の分析を行った．はじめに質感データセットを作成し，このデータセットを基に，分類タスクと生成タスクを設計し，LVLM の評価を実施した．評価の結果，分類タスクの正解率が高い LVLM は，生成タスクにおいても高いスコアを示す可能性が示唆された．またこの結果から，我々が提案した分類タスクは，LVLM の質感知覚能力の評価だけでなく，人間知覚の整合性まで簡易に評価できる可能性がある．今後の展望として，より精緻に分類タスクと生成タスクの相関関係を調査したい．

謝辞

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．特に，アノテーターとして参加していただいた Tohoku NLP グループのメンバーの佐藤魁さん，岩川光一さんにはこの場を借りて御礼申し上げます．本研究は，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research)，および，文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです．本研究成果（の一部）は，データ活用社会創成プラットフォーム mdx を利用して得られたものです．

参考文献

[1] 西田眞也. 「多元質感知」における質感研究. 日本画像学会誌, Vol. 57, No. 2, pp. 189–196, 2018.

[2] 上村卓也, 澤山正貴, 西田眞也. 多様な質感認識の情報処理に用いられる画像特徴を統一的に説明するためのニューラルネットワークモデルの検討. 人工知能学会全国大会論文集, Vol. JSAI2018, pp. 4O1OS3a02–4O1OS3a02, 2018.

[3] Waka Fujisaki. Multisensory Shitsukan perception. **Acoustical Science and Technology**, Vol. 41, No. 1, pp. 189–195, 2020.

[4] Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In **Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)**, pp. 109–132, 2021.

[5] Roland W. Fleming. Material Perception. **Annual Review of Vision Science**, Vol. 3, No. Volume 3, 2017, pp. 365–388, 2017.

[6] Hidehiko Komatsu and Naokazu Goda. Neural Mechanisms of Material Perception: Quest on Shitsukan. **Neuroscience**, Vol. 392, pp. 329–347, 2018.

[7] Charles Spence. Shitsukan — the Multisensory Perception of Quality. **Multisensory Research**, Vol. 33, No. 7, pp. 737–775, 2020.

[8] Bei Xiao and William Kistler. Perceptual Dimensions of Material Properties of Fabrics in Dynamic Scenes. **Journal of Vision**, Vol. 15, No. 12, pp. 938–938, 2015.

[9] 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**, pp. 740–755, 2014.

[10] 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**, cs.CL/2303.08774v6, 2023.

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. **arXiv preprint**, cs.CL/2407.21783v3, 2024.

[12] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. **arXiv preprint**, cs.CL/2409.12191v2, 2024.

[13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. **arXiv preprint**, cs.CL/2408.03326v3, 2024.

[14] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024.

[15] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? **arXiv preprint**, cs.CL/2405.02246v1, 2024.

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. **arXiv preprint**, cs.CL/2310.03744v2, 2023.