Vision Language Model が持つ画像批評能力の評価手法の提案

† 齊藤成輝 [1]　† 林和樹 [2]　† 井手佑翼 [2]　坂井優介 [2]
鈴木刀磨 [2]　郷原聖士 [2]　大西雄真 [1]
‡ 上垣外英剛 [2]　林克彦 [1]　渡辺太郎 [2]
[1] 北海道大学　[2] 奈良先端科学技術大学院大学
‡ kamigaito.h@is.naist.jp

概要

Large-scale vision language models（LVLMs）は，画像を基に文章を生成する LLM である．本研究では LVLMs の応用として，画像の良い点や悪い点を批評するタスクを考えるが，LVLMs がそのような能力をどの程度備えているのかは明らかとなっていない．そのため，本研究では LVLMs の画像批評能力を評価する手法を提案し，人間のアノテータによって付けられた批評文の優劣と LVLMs によって判断される批評文の優劣にどの程度の相関が見られるか調査した．提案手法で作られた評価用データセットに対して，一般に性能が高いとされている LVLM ほど人間との間に正の相関が見られ，評価手法として一定の妥当性が確認された．

1　はじめに

大規模言語モデル（LLMs; Large Language Models）[1] は，視覚基盤モデル [2] との統合により，入力された画像を基に文章を生成する能力を獲得した．このように視覚と言語の融合（V&L）を実現した Large-scale vision language models（LVLMs）[3, 4, 5, 6, 7, 8, 9] は，V&L のベンチマークにおいて大きな成功を収めており，更なる応用への可能性が期待されている．

こうした応用の可能性の一つとして，LVLM を使った画像への批評が挙げられる．例えば，写真や絵画などの画像コンテストでは専門家の批評に基づいて優れた画像が選ばれるが，批評は高度で複雑な作業となる．そのため，LVLM を用いた画像批評を実現できれば，画像コンテストを自動化し，それを日常的に開催することも可能となる．画像作品に対する批評が気軽に得られることで，ユーザは作品の良い点や悪い点を簡単に知ることができ，画像制作技術の向上にも繋がる．

図 1　批評文の例：画像の情報と整合性が取れており，できる限り客観的な視点から良い点と悪い点に言及．

画像への批評では，基本的に，画像の内容と一致する情報を捉えた上で，その良い点と悪い点をできる限り客観的な視点から説明する必要がある（図 1）．従って，まずは，これら二つの基本的な能力

- 画像の内容と整合性の取れた情報を捉える能力
- できる限り客観的な視点で批評する能力

を考えることで，LVLMs が画像の批評能力をどの程度備えているのか評価する．本研究で提案する手法は，同一の画像に対して，強力な LVLM（本研究では GPT-4V を使用）によって生成された 5 種類の批評文を，画像との整合性，客観性，文章の流暢性に基づいて人間のアノテータが順位付けする．また，他の LVLM を用いて，画像・批評文生成用 prompt・生成された批評文の三つ組から Perplexity を計算し，それに基づいて，5 種類の批評文の順位を計算する．この結果と人間がアノテーションした結果の順位相関からモデルの性能を評価する．

実験では，様々なジャンルの良質な画像が掲載されている英語版 Wikipedia の「Featured pictures」項目から 207 枚の画像を取得した．取得した画像と，それに対応する 5 種類の批評文，及び人間による順位付け評価結果を，LVLM の批評能力の評価用データセットとする．データセットは英語と日本語版の 2 種類を作成した．アノテーション結果から日英共

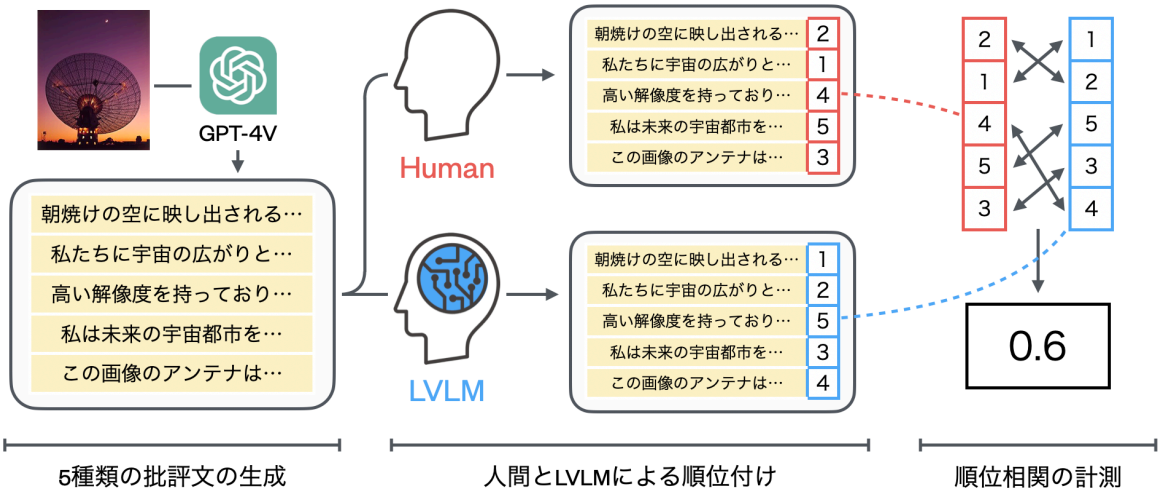

図 2　提案手法の概略：LVLM が持つ画像の批評能力を評価するための手法.

に，100 枚を超える画像に対して，2 名のアノテータ間で平均として 0.7 を超える強い相関が確認された．また，評価用データセット上で，三つの LVLMs に対して，Perplexity に基づく順位を計算し，人間のアノテーション結果と順位相関を計測したところ，一般に性能が高いとされている LVLM ほど人間との間に正の相関があることが明らかとなった．

データセットや各種プログラムは公開しており[1)]，今後更なる発展が期待される LVLMs の画像批評能力の検証などに利用できる．

2　関連研究

2.1　LVLM

近年，LLM は顕著な成功を収めており，言語だけでなく画像や音楽など，多様な入力形式との統合が進んでいる．本研究では，特に画像入力を統合した LLM である LVLM に焦点を当てる．初期段階の研究である Flamingo [3] は，現在の基本となる Visual Encoder と LLM の結合の可能性を示した．これは少数またはゼロショットでの Visual Question Answering（VQA）において優れた性能を発揮した．後続の研究では，画像を用いた Instruction Tuning で少量のデータでも高い性能を示した LLaVa [5, 6]，Visual Encoder を再学習して画像との整合性を高めた mPLUG-Owl [7, 8]，位置関係を認識する Vision Language Adapter を組み込み，キャプション生成など幅広い視覚中心の V&L タスクにおいて最高精度を達成した Qwen-VL [9] などが提案されてきた．

2.2　LVLM の評価

初期段階の LVLM の研究では，多様なサブタスク [10, 11] における評価が主に行われていた．しかし，LVLM の進展に伴ってその応用範囲は広がり，LVLM が持つ知覚や認知などの高度な能力に焦点を当てた手法での評価へと遷移した [12, 13, 14]．最近では，評価対象の LVLM が互いに評価し合う手法や，評価者や評価の補助として強力な LVLM（例えば GPT-4V）を活用する手法など，LVLM 自体を活用する手法が増えている．本研究のように，データ作成に GPT-4V を利用する方法も模索されている [15, 16, 17, 13, 14].

3　提案手法

本節では LVLMs が持つ批評能力を評価するための提案手法について説明する．提案手法は，図 2 に示す通り，以下の 3 つのステップから構成される．

- 5 種類の批評文の生成
- 人間と LVLM による順位付け
- 順位相関の計測

次節以降では，各ステップの詳細を説明する．

3.1　5 種類の批評文の生成

まず，画像を複数ジャンルからバランス良く収集し，各画像ごとに 5 種類の批評文を生成する．

画像の取得　英語版 Wikipedia の「Featured pictures」項目から画像を取得する．「Featured pictures」項目は，Wikipedia に掲載されている写真，イラスト，図表の中で，特に優れたものをユーザの投票に

1)　https://github.com/naist-nlp/Hackathon-2023-Summer

より選出した項目である．この項目には，美術作品，自然景観，歴史的出来事，科学など多様なジャンルの良質な画像が掲載されている．このように，画像コンテストと類する場面で選出され，主題が明瞭で良質な画像で構成されている特徴を踏まえ，「Featured pictures」項目を画像の取得元に選んだ．

批評文の生成　批評文の生成には，GPT-4V を使用する．研究目的を考慮すると，批評対象となる幅広い種類の画像に対して，深い知識を持つ専門家が，批評文を人手で作成することが理想的である．しかし，実際には，そのような幅広い専門家を集めることは現実的ではない．また，非専門家が外部知識を参照しつつ批評文を作成する方法も考えられるが，これには膨大な時間と労力が必要となる．

我々の事前調査から，GPT-4V により生成される批評文は，非専門家が作成するものと比較して高品質であり，研究目的を満たすのに十分な品質であることが確認された．この調査結果に基づき，本研究では GPT-4V を使用して批評文を生成した．

GPT-4V で簡易に温度パラメータを変化させて異なる批評文を生成した場合，文章の品質にほとんど差がなく，人間による順位付けが困難となる．そのため，下記に示すように prompt を工夫して，異なる 5 種類の批評文を生成する．prompt の赤字の箇所で

> Please describe five different review texts about the good points and room for improvement of the image, following the constraints below:
> 1. Each review text should have different content.
> 2. The length of each review text should be almost the same.
> 3. Do not include bullet points within the review texts.
> 4. The review texts should be described in the following order: "Objective and reasonable," "Subjective but reasonable," "Objective but unreasonable," "Subjective and unreasonable," and "Subjective and containing an error."
> 5. Each review text should describe both the good points and room for improvement of the image.
> 6. If the image has no room for improvement, explicitly state that within the review text.

は，整合性（reasonable で誤りを極力含まない）・客観性（Objective）という点を重視し，5 種類の指定を記述した．整合性が reasonable となっているのは，prompt エンジニアリングの結果である．一般に，批評は主観性を伴うものであるが，画像コンテスト等での利用を考えると，整合性を前提として，客観性をできるだけ重視する必要がある．そのため，生成される 5 種類の批評文は前から順に質の高いものが生成されることを仮定する．実験では，この順位（prompt 順位）とアノテートされた順位との相関を計測することで，この仮定の妥当性を裏付ける．

また，生成した批評文には「Note: This review contains an error as the stars are not blurred in the image provided」のような批評文自体を否定する一見矛盾した表現が末尾に含まれる場合がある．これは ChatGPT が強化学習による人間のフィードバック（RLHF）[18] に基づいて出力を制御しているために生じる現象と考えられる．このような表現は，人手によるチェックを行い，該当する文を削除した．

3.2　順位付け

次に，取得した各画像に対する 5 種類の批評文をアノテータと LVLM によって順位付けする．

アノテータ　本研究では日本語と英語の批評文を作成する．順位付け作業を行う前に，下記に示される詳細な説明と指示を行い，評価の一貫性を確保した．以下は英語の説明と指示であり，日本語データのアノテータにはこれを日本語にしたものを与える（付録 A.1 参照）．概要として，画像に含まれる情報

> Below are the images and their review texts. Please rank the review text of each image from 1 to 5, in order of appropriateness. Please note that the numbers from 1 to 5 are not scores but rankings, and the smaller the number, the more appropriate it is. There should be no ties, and each rank from 1 to 5 should always appear once.
>
> Please judge the appropriateness by the following aspects in the following order. That is, first, rank the texts by truthfulness. If there are equally truthful texts, rank them by consistency. Similarly, if they are equal also in consistency, rank them by informativeness; if they are equal also in it, rank them by objectivity; if they are equal also in it, rank them by fluency.
>
> 1. Truthfulness: Is it free of false information?
> 2. Consistency: Does it correspond to the image?
> 3. Informativeness: Does it describe detailed information or features of the image?
> 4. Objectivity: Is it an objective description?
> 5. Fluency: Is it grammatically correct?
>
> If the text contains unfamiliar information, you may use a dictionary or search engine. However, please do not use a generative AI such as ChatGPT or image search.

と整合性があり，客観的で流暢性の高い批評を重視する指示となっている．5 つの批評文は，提示する順序から質を推測できてしまう状況を避けるため，シャッフルしたうえでアノテータに提示する．

LVLM　LVLM による順位付けには，Perplexity を使用する．Perplexity は，あるサンプルの文章がモデルに与えられたとき，モデルがその文章をどの程度予測できるかを示し，サンプルの文章に対するモデルの不確実性を表す．本研究では，「Please describe a review text about the good points and room for improvement of the image」という prompt，画像及び批評文を，評価対象の LVLM に与え Perplexity を計測する．ここで，Perplexity が低いほど，LVLM の鑑賞文に対する不確実性が低い，つまり，その LVLM

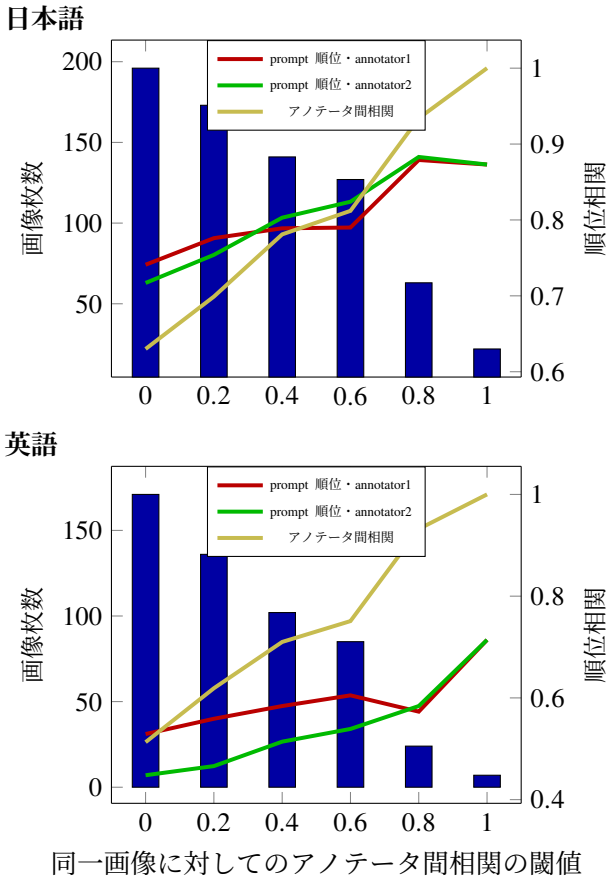

図 3　画像枚数と各種相関の変化: 棒グラフは閾値を超える画像枚数，折れ線グラフは各種相関を表す．

にとって予測可能性が高い批評文と考えられる．

3.3　順位相関の計測

最後に，人間のアノテータと LVLM の順位付けの結果から，両者の順位相関を測定し，平均する．この測定には，スピアマンの順位相関係数を使用する．この係数は，−1 から 1 の範囲の値をとり，−1 は完全な逆順を，1 は完全な同順を示す．順位相関が高い場合，つまり係数が 1 に近い場合，人間のアノテータが「優れている」と判断した批評文は，LVLM にとって予測可能性が高く，また「劣っている」と判断した批評文は，予測可能性が低いと言える．従って，人間との順位相関が高い LVLM ほど，批評文の評価能力は高いと考える．本研究では，各画像について測定した順位相関の平均を報告する．

4　検証

画像データの取得　本研究では，Wikipedia の「Featured pictures」における 15 のジャンルから合計で 207 枚の画像を取得した．ジャンルは Animals, Artwork など多岐に渡る（詳細は付録 A.2 を参照）．

	mPLUG-Owl	mPLUG-Owl2	Qwen-VL
日本語	0.135	0.441	**0.510**
英語	0.261	0.325	**0.417**

表 1　LVLM・アノテータ間の相関．

人手アノテーション　日本語データへのアノテーションは，2 名の日本語母語話者が行った．英語データのアノテーションは，ネイティブ及びそれに準ずる英語リテラシを持つアノテータ 2 名を採用した．人手によるアノテーションの結果，一部の画像ではアノテータ間で評価が大きくバラついていることが確認された．そのため，アノテータ間の相関に閾値を設けて，画像のフィルタを行った．図 3 は残る画像枚数・prompt 順位と各アノテータの相関・アノテータ間の相関の三つを示している．この結果より，閾値を 0.4 とした場合，日・英両方で平均として 0.7 を超える強い相関が確認され，更に，100 枚以上の画像が基準を満たすことが確認できた．

また，整合性及び客観性を重視する批評文生成において，本研究の prompt では，前から順に質の高いものが生成されることを仮定していたが，図 3 に示される prompt 順位と各アノテータの強い正の相関より，この仮定にはある程度の妥当性があることが確認された．

結果　評価対象の LVLMs としては，mPLUG-Owl [7]，mPLUG-Owl2 [8]，Qwen-VL [8] の 3 つを採用した．LVLM とアノテータの間で順位付けがどれほど一致するか測るため，順位相関係数を計算した．各 LVLM について，2 人のアノテータそれぞれとの相関を計算し，その平均を LVLM・アノテータ間相関とした．結果を，表 1 に示す．検証の結果，日英両言語で，Qwen-VL，mPLUG-Owl2，mPLUG-Owl の順にアノテータとの相関が高くなった．

5　まとめ

本研究では，LVLMs の応用として画像批評タスクを考え，基本的な批評能力と考える，画像と整合性のある情報を捉える能力，できる限り客観的な視点から分析する能力を明らかにするための新しい手法を提案する．3 種類の LVLMs にて検証した結果，一般に性能が高いとされている LVLM ほど人間との間に正の相関が見られ，評価手法として一定の妥当性が確認された．本研究で作成した評価用のデータセットは公開予定である．

謝辞

本研究は JSPS 科研費 JP23H03458 の助成を受けたものです.

参考文献

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[2] 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.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 23716–23736, 2022.

[4] 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.

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

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

[7] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. **arXiv preprint arXiv:2304.14178**, 2023.

[8] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. **arXiv preprint arXiv:2311.04257**, 2023.

[9] 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.

[10] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In **International Conference on Computer Vision (ICCV)**, 2015.

[11] 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.

[12] 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.

[13] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. **arXiv preprint arXiv:2307.16125**, 2023.

[14] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models, 2023.

[15] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. **arXiv preprint arXiv:2306.09265**, 2023.

[16] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? **arXiv preprint arXiv:2307.06281**, 2023.

[17] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. **arXiv preprint arXiv:2308.06595**, 2023.

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.

A　参考情報

A.1　日本語データの作成

日本語データの作成に使用した prompt と，アノテータへの指示文を以下に載せる．

画像の良い点と改善可能な点について記述した異なる5つの鑑賞文を
下記の制約に従って記述してください。
1. 各鑑賞文の内容が異なっている。
2. 各鑑賞文の長さがほぼ同じである。
3. 鑑賞文中に箇条書きを含めない。
4. 鑑賞文は「客観的で整合性が高い」、「主観的だが整合性が高い」、
「客観的だが整合性が低い」、「主観的で整合性が低い」、
「主観的で誤りを含む」ものをこの順序で5つ記述する。
5. 各鑑賞文が画像の良い点と改善可能な点について記述している。
6. 画像に改善可能な点が存在しない場合は鑑賞文中にそのことを明示する。

以下は画像とその鑑賞文です。各画像の鑑賞文に、適切であると思う順に1から5までの順位をつけてください。ただし、1から5までの数値はスコアではなく順位であり、小さいほど適切であることを表す点に注意してください。同率の順位はないものとし、1から5までの順位を必ず一度ずつ使ってください。

適切さの判断にあたっては、次の観点に次の順で注意してください。すなわち、まず忠実性に基づいてテキストに順位をつけてください。忠実性が同等のテキストがあった場合は、一貫性に基づいてそれらの順位を決めてください。以下同様に、一貫性も同等なら情報性、それも同等なら客観性、それも同等なら流暢性に基づいて、順位をつけてください。

1. 忠実性：嘘が含まれていないか
2. 一貫性：画像に対応した説明がなされているか
3. 情報性：画像の詳細な情報(特徴)が記述されているか
4. 客観性：客観的な説明がなされているか
5. 流暢性：文法に破綻がないか

テキスト中に知らない事柄が含まれていた場合は、辞書や検索エンジンを使用しても構いません。ただし、ChatGPTなどの生成AIや、画像検索は利用しないでください。

以下は画像とその鑑賞文です。各画像の鑑賞文に、適切であると思う順に1から5までの順位をつけてください。ただし、1から5までの数値はスコアではなく順位であり、小さいほど適切であることを表す点に注意してください。同率の順位はないものとし、1から5までの順位を必ず一度ずつ使ってください。

適切さの判断にあたっては、次の観点に次の順で注意してください。すなわち、まず忠実性に基づいてテキストに順位をつけてください。忠実性が同等のテキストがあった場合は、一貫性に基づいてそれらの順位を決めてください。以下同様に、一貫性も同等なら情報性、それも同等なら客観性、それも同等なら流暢性に基づいて、順位をつけてください。

1. 忠実性：嘘が含まれていないか
2. 一貫性：画像に対応した説明がなされているか
3. 情報性：画像の詳細な情報(特徴)が記述されているか
4. 客観性：客観的な説明がなされているか
5. 流暢性：文法に破綻がないか

返答には、理由を含めず、順位だけ出力してください。
返答は以下の形式でお願いします。
文1:2位,文2:3位,文3:1位,文4:5位,文5:4位

A.2　画像データのジャンル内訳

下記に取得した画像データにおける 15 のジャンルとその内訳を示す．( ) 内は取得した画像数を表す．

Animals (17) / Artwork (17) / Culture, entertainment, and lifestyle (16) / Currency (15) / Diagrams, drawings, and maps (15) / Engineering and technology (17) / Natural phenomena (15) / People (14) / Places (17) / Plants (16) / Sciences (15) / Space (15) / Vehicles (5) / Other lifeforms (3) / Other (10)

A.3　GPT-4V の評価

GPT-4V に対しても順位付けによる評価を行った．なお，GPT-4V は Perplexity を算出できないため，アノテータと同様に，画像・指示文・批評文の三つ組みから順位付けを行った．順位付けの結果から人間のアノテータとの順位相関係数を算出し，同一画像に対するアノテータ間相関に閾値を設け，その値に対応する順位相関係数の結果を図 4 にまとめた．与えた指示文は以下に載せる．

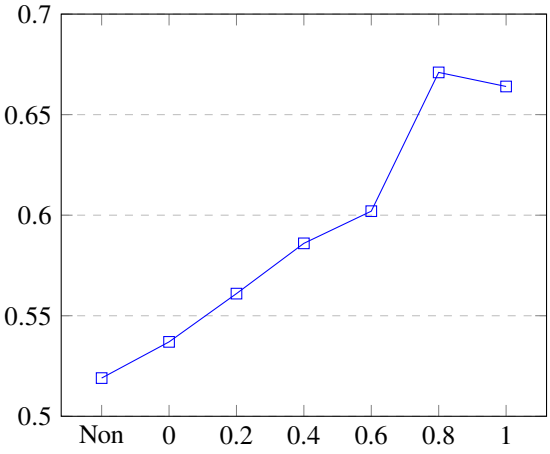

図 4　GPT-4V の評価．