Vision-Language Model の内部確信推定指標の提案

宮田 侑佳 [1] 西潟 優羽 [1] 水沼 千枝 [2] 奥村 紀之 [3] 倉光 君郎 [4]
[1] 日本女子大学大学院 理学研究科 [2] 日本女子大学 家政学部
[3] 武庫川女子大学 社会情報学部 [4] 日本女子大学 理学部
m2116081my@ug.jwu.ac.jp kuramitsuk@fc.jwu.ac.jp

概要

本稿では、Vision Language Model (VLM) の出力が画像に捉えられている内容を根拠としているか推定する手法を提案する。VLM は画像とテキストを入力として受け取るが、内部的には言語モデルが支配的であり、画像を十分に活用せず言語知識の影響が強く表れる傾向がある。この問題を定量的に評価するため、本研究では、画像の有無による内部確信の変化を比較することで、VLM がその出力において、画像を活用しているかを定量的に推定する 2 つの指標を考案した。指標 I は、Perplexity (PPL) の原理に基づき、出力確率分布から内部確信を推定する新規手法である。指標 II は、既存研究で提案された Self-Consistency (SC) の原理を応用し、複数回の試行から一貫性を測定することで指標 I の妥当性を補完する。画像の有無による VLM の振る舞いの変化を利用し、提案指標が VLM の内部確信を推定する指標になる可能性を確認した。

1 はじめに

VLM は、画像とテキストを統合的に処理し、視覚的質問応答や画像説明生成などのタスクにおいて優れた性能を示しており、実社会での応用も進んでいる。しかし、VLM の実用化に伴い、その信頼性に関する懸念も高まっている。企業が AI 導入において最も頻繁に経験するリスクとして「不正確さ」が報告されており、説明可能性などの信頼性に関わる課題が上位を占める [1]。

実社会のアプリケーションに VLM が導入される際、その出力が画像に根拠を持つかを評価することは、タスクのパフォーマンスを測定することと同様に不可欠と言える [2]。本研究の主要な問題は、VLM の出力が画像を十分に活用せず、言語知識から受ける影響が大きい可能性を否定できないことにある。

VLM は画像とテキストの両方に基づいて出力を生成することが期待されるが、実際には、その出力がテキストプロンプトや事前学習で獲得した知識に強く影響される場合がある。近年の研究により、VLM が画像の変更を見落とし、代わりに知識に基づいて誤答してしまう失敗は画像情報の不足ではなく、記憶された知識が視覚的分析を上書きすることに起因すると指摘される [3, 4]。

そこで本稿では、VLM の信頼性評価における新たな検証アプローチとして内部確信を推定する指標 I、II を提案する。画像の有無による VLM の出力挙動に着目し、内部確信の変化量を算出することで VLM が画像を活用しているかを定量化する。

実験では、着物画像、オープンモデルを対象とし、提案する両指標を扱った。プロンプト設計により、両指標間には相関が見られ、PPL の原理に基づく Δc_M が内部確信を捉える指標になる可能性が示された。

2 内部確信

本研究における内部確信とは、VLM が自身の出力に対してどの程度確信を持っているかを示す指標である。VLM の最終的な出力テキストのみを見ても、その決定が画像に基づいているのか、事前知識によるものなのか、それともテキストプロンプトに駆動されているのか、これらは判断されにくい。そのため、VLM の出力に対するプロセスを推定するには、その内部状態に着目した指標が必要となる。

2.1 提案手法の概要

本稿では、画像の有無による内部確信の変化を比較することで、VLM の出力が画像を活用しているかを定量的に推定する 2 つの指標を提案する。

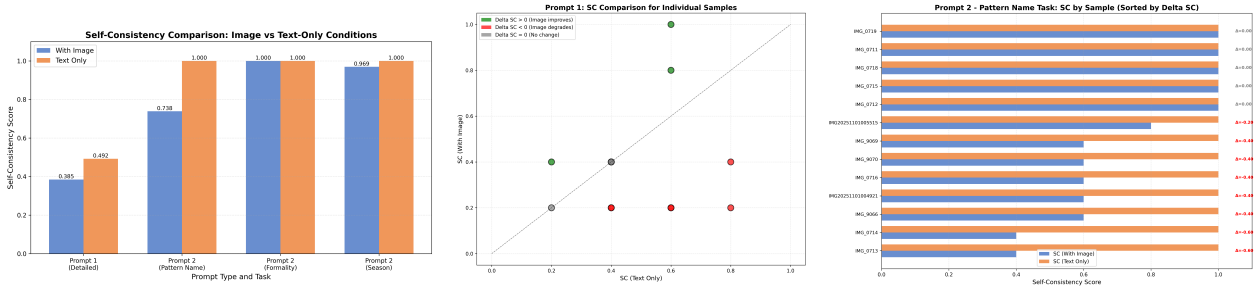

図 1: プロンプト形式別の Self-Consistency（SC）比較と詳細分析

2.2 指標 I：PPL ベースの内部確信

指標 I は、PPL の原理に基づく。PPL は言語モデルの評価において一般的に使用される指標の一つである。言語モデルの PPL は式 (1) で定義される。

$$\mathrm{PPL} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log p_t\right) \tag{1}$$

ここで、$p_t = p(y_t|y_{<t}, x)$ は、出力系列 $y_{1:T}$ における各トークン y_t の条件付き確率である。PPL はモデルが次の単語を予測する際の不確実性を表す。
この定義により、以下の関係が成立する。

$$C = \text{internal confidence} = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\log p_t\right) = \frac{1}{\mathrm{PPL}} \tag{2}$$

つまり、内部確信が高いほど C の値が高くなり、内部確信度が低いほど C の値が小さくなることを意味する。この関係により、モデルの確信度を直感的に解釈可能な指標として扱うことができる。

画像有無による内部確信の差分測定

本研究の提案は、**画像ありの場合の内部確信度 C_n^{img} と画像なしの場合の内部確信度 $C_n^{no\_img}$ を測定し、その差分を計算する**ことである。

N 個のサンプルに対して、画像有無による内部確信の平均差分 Δc_M を以下のように定義する。

$$\Delta c_M = \frac{1}{N}\sum_{n=1}^{N}(C_n^{img} - C_n^{no\_img}) \tag{3}$$

- C_n^{img}: n 番目のサンプルにおける画像ありの場合の内部確信
- $C_n^{no\_img}$: n 番目のサンプルにおける画像なしの場合の内部確信
- N: 評価に用いるサンプル数

指標の解釈 Δc_M の値から、以下のように判断する。

- $\Delta c_M \neq 0$ ($\Delta c_M > 0$ または $\Delta c_M < 0$): 画像の有無で内部確信が変化しており、VLM が画像を処理に活用している可能性が高い
- $\Delta c_M \approx 0$: 画像の有無で内部確信が変わらず、VLM が画像を無視している可能性が高い

Δc_M の符号（正か負か）それ自体は、視覚情報の活用度を直接示すものではない。重要なのは、画像の存在がモデルの内部状態に影響を与えていることを意味し、画像が何らかの形で処理・活用されていることを示唆する。一方、$\Delta c_M \approx 0$ の場合、画像の有無がモデルの内部確信に影響を与えていない可能性を示す。

2.3 指標 II：SC ベースの内部確信

Wang et al.[5] は、言語モデルの推論における一貫性評価手法として Self-Consistency（SC）を提案した。SC の特徴は、推論経路の違いを無視し、最終的な答えの意味での一貫性を評価するという点である。具体的には、同一入力に対してモデルが生成した複数の出力から、最頻回答を選択することで、意味レベルでの一貫性を評価する。

$$a^* = \arg\max_a \sum_{i=1}^{m} \mathbf{1}(a_i = a) \tag{4}$$

ここで、m はサンプリング回数、a_i は i 回目の生成結果における最終回答である。Yona et al.[6] は、先行研究 [7, 8, 9] に従い、Wang et al. が提案した SC の概念を拡張し、これを内部確信として明示的に定式化した。彼らは、同一入力を複数回生成した際の回答間の一致度によって、モデルの内部確信を測定できることを示した。

本研究では、Yona et al.[6] が提案した SC に基づく内部の概念を、VLM による出力の画像活用に関する評価に拡張している。従来の言語モデル研究では、内部確信は主にテキスト生成の一貫性を評価するために用いられてきた。本研究では、この概念を

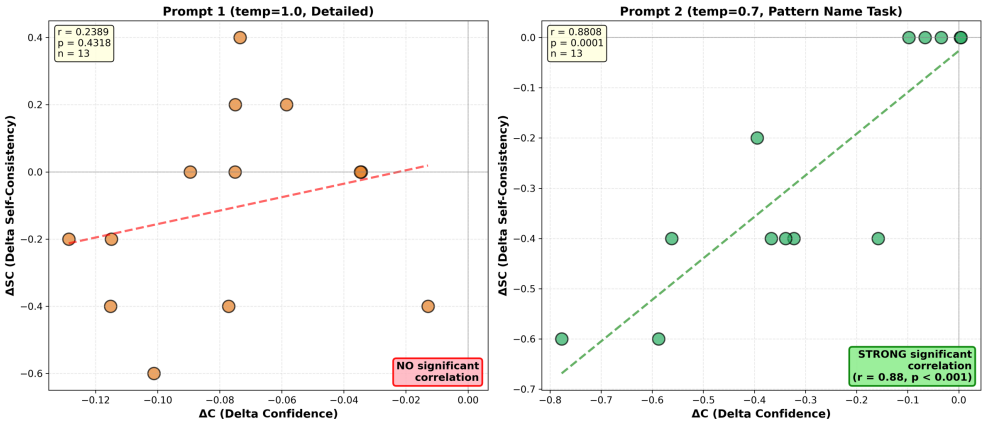

図 2: プロンプト別の推定指標 IΔc_M と指標 IIΔSC の相関関係

マルチモーダル設定に適用し、画像の有無による内部確信の変化を測定することで、VLM が画像情報をどの程度活用しているかを評価する。

指標 II は、**出力の意味レベルでの一貫性に基づく確信**である。同一入力を複数回繰り返した際の出力の安定性を、SC によって測定する。

本研究では、SC を最頻回答が現れた回数を総繰り返し回数 K で割ったものとして定義する。

$$SC = \frac{1}{K} \max_{a} \sum_{i=1}^{K} \mathbf{1}(a_i = a) \tag{5}$$

ここで、a_i は i 回目の生成における最終回答である。

これは、モデルの「矛盾しない割合」に相当し、**内部確信を直接的に推定する指標**となる。よって、SC が 1 に近い場合、モデルは一貫した出力を生成しているため、高い確信を示す。一方、SC が 0 に近い場合、出力は多様化しているため、モデルの不確実性を示す。

画像有無による SC の差分測定

画像あり/なしで、モデルがどれだけ一貫した回答を維持できるかを比較し、その変化量 ΔSC を VLM の処理過程における画像利用の指標として用いる。画像ありの場合の SC（SC_{img}）と画像なしの場合の SC（SC_{no_img}）の差分 ΔSC は，式 (6) に示すように定義される。

$$\Delta SC = SC_{img} - SC_{no\_img} \tag{6}$$

指標の解釈

ΔSC の値により、以下のように解釈できる。

- $\Delta SC \approx 0$: 画像有無で内部確信が類似しており、画像が出力にほとんど影響を与えていない
  - VLM が画像を利用していない可能性が高い
- $\Delta SC \neq 0$ ($\Delta SC > 0$ または $\Delta SC < 0$): 画像の有無で内部確信が変化している
  - VLM が画像を利用している可能性が高い

これより、VLM が処理において画像を利用しているか、出力の一貫性という観点から評価する。

3 実験

本実験では、提案した Δc_M が、VLM の内部確信を推定する有効な指標であるかを検証する。

本稿では、画像の有無による VLM の振る舞いの変化を利用する。具体的には、同一の質問に対して、以下の 2 つの条件で VLM の応答を比較する：

- **画像あり条件**: 質問と画像の両方を入力
- **画像なし条件**: 質問のみを入力（テキストのみ）

この 2 条件間での挙動の変化を観察することで、画像情報が VLM の内部状態（確信度・一貫性）に与える影響を定量化する。変化量が大きい場合、VLM が画像情報を活用して内部確信を形成していることを示す。

データセット・モデル

着物画像を使用した。各画像には、桜、牡丹、扇、竹、鳥などの多様な柄が含まれる。本データセットは、被服学の専門家との共同研究における着物文化継承支援プロジェクトの一環として収集された。本プロジェクトは、着物の柄や色に内包された意味を体系的に整理することで、着物文化の理解を促進することを目的としている。モデルは、Qwen3-VL-8B-Instruct[1)] を使用した。

プロンプト設計

異なるタスク形式が VLM の内部確信にどのように影響するかを調査するため、2 種類のプロンプト

1) 確率分布へのアクセスが可能なマルチモーダルモデル

表 1: Qwen におけるプロンプト設計別の内部確信推定指標（C^{img}，C^{no_img}，ΔC）の比較

指標	Prompt1（詳細説明）	Prompt2（1 語）
サンプル数	13	39
平均 ΔC	−0.076	+0.044
ΔC 中央値	−0.075	+0.00006
ΔC 傾向	負（100%）	正負混在（正 51%／負 49%）
標準偏差（ΔC）	0.035	0.357
IQR	0.043	0.437
統計的有意性	$p < 0.001$	有意差なし
効果量（Cohen’s d）	−2.20（大）	小（$\lvert d \rvert \ll 0.5$）
C^{no_img}	≈0.82（固定）	≈0.99（ほぼ固定）
C^{img}	0.69–0.80	0.67–1.00
画像なしの応答	定型的（「なし」）	高確信の定型回答が多い
画像ありの応答	確信が低下	確信が上下

を設計した。

プロンプト 1：詳細説明型（多項目回答形式）

複数の質問項目を含む包括的なプロンプトである。モデルに対して、柄の名称、意味、適した場面、避けるべき場面の 4 点を順番に回答するよう求める。temperature はデフォルト値 (0.7) を使用した。

- **回答形式**: 自由記述形式（番号付き多項目回答）
- **特徴**: モデルの推論能力を幅広く評価

プロンプト 2：単一質問型（1 語回答形式）

各質問を個別に提示し、1 語での簡潔な回答を求めるプロンプトである。各質問には「分からない場合は『不明』とだけ答えてください」という指示を含め、モデルが不確実性を明示的に表現できるようにした。temperature は同様に、0.7 を設定した。

- **回答形式**: 1 語回答（制約的）
- **特徴**: 回答の揺らぎを抑制して評価

プロンプト 2 は、より決定論的な回答を促したため、一貫性の高い結果が期待される。

3.1 実験結果

提案する推定指標 I の振る舞い

表 1 に，Q2 種類のプロンプト設計による提案指標 I による内部確信の変化を集約する。

プロンプト 1 では、画像入力によって出力の確信が一貫して低下した。ΔC は、VLM の出力が画像を根拠にしているかを捉える指標であると示唆している。一方、プロンプト 2 では、画像入力が出力挙動に与える影響は安定して観測されなかった。

これらの結果は、VLM の出力挙動がプロンプト設計によって画像入力への応答性が変化し、Δc_M はその変化を定量的に捉える指標として機能する可能性も示している。

提案する推定指標 II の振る舞い

図 1 に、プロンプト形式別の SC 分析を示す。左図より、プロンプト 2 では画像なし条件で SC=1.0 となり、「不明」と一貫して回答したことが分かる。中図では、プロンプト 1 において画像情報により一貫性が低下したサンプル（赤色）と向上したサンプル（緑色）が混在している。右図より、プロンプト 2 の柄の名称タスクでは、画像なし条件で全サンプル SC=1.0、画像あり条件では SC=0.4-1.0 と分布していることが示された。これらの結果は、プロンプト設計が SC に大きく影響することを示している。

3.2 Δ C と Δ SC の相関分析：提案指標の妥当性検証

図 2 は、提案する推定指標 I,Δc_M と指標 II,ΔSC の関係を比較したものである。

詳細説明型では、Pearson 相関係数 $r = 0.24$（$p = 0.43$）であり、統計的に有意な相関は観察されなかった。データポイントは散在しており、回帰直線の傾きも緩やかである。これは、多項目自由記述形式というプロンプトの性質により、モデルの内部確信と出力の一貫性が乖離したためと考えられる。

この結果は、単にプロンプト 1 では指標が機能しないことを示すのではなく、両指標が測定する内部確信を推定するには、プロンプト設計、パラメータ設定に対して指標の相性について、考慮する必要があることを留めておきたい。

一方で、プロンプト 2 の柄の名称タスクでは、**Pearson 相関係数 $r = 0.88$（$p < 0.001$）という極めて強い正の相関**が観察された。データポイントは回帰直線に沿って密に分布しており、決定係数 $R^2 = 0.77$ は、Δc_M が ΔSC の変動の 77%を説明していることを意味する。両指標の強い相関は、それぞれが異なる方法で同じ内部確信という現象を捉えている可能性を示唆した。

4 おわりに

本稿では、内部確信を推定する ΔC を提案し、既存指標を応用した ΔSC との関係を検証した。実験により、両指標に強い相関が観察された。これは、VLM の信頼性評価における新たな検証アプローチの可能性を示す。

ただし、本研究は単一モデルでの検証にとどまっており、提案手法の汎用性を確立するには、他のモデルでの検証が不可欠である。また、VLM が画像情報と事前知識をどのように統合して内部確信を形成するのか、そのメカニズムに対する検証も重要な課題として残されている。

参考文献

[1] Alex Singla, Alexander Sukharevsky, Lareina Yee, Michael Chui, Bryce Hall, and Tara Balakrishnan. The state of ai in 2025: Agents, innovation, and transformation. Technical report, McKinsey & Company / QuantumBlack, AI by McKinsey, November 2025. Accessed: 2025-11-05.

[2] Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. **arXiv preprint arXiv:2505.20236**, 2025.

[3] An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. **arXiv preprint arXiv:2505.23941**, 2025.

[4] Jiachen Yu, Yufei Zhan, Ziheng Wu, Yousong Zhu, Jinqiao Wang, and Minghui Qiu. Vfaith: Do large multimodal models really reason on seen images rather than previous memories? **arXiv preprint arXiv:2506.11571**, 2025.

[5] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. **arXiv preprint arXiv:2203.11171**, 2022.

[6] Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? **arXiv preprint arXiv:2405.16908**, 2024.

[7] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. **arXiv preprint arXiv:2302.09664**, 2023.

[8] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In **Proceedings of the 2023 conference on empirical methods in natural language processing**, pp. 9004–9017, 2023.

[9] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. Fine-tuning language models for factuality. In **The Twelfth International Conference on Learning Representations**, 2023.