マルチモーダル大規模言語モデルにおける工業製品画像の認識性能調査

遠藤隆夫[1] 岡部健太[1] 石上将太郎[1] 中村光貴[1]
仁平雅也[1] 乙村浩太郎[1] 羽藤淳平[1]
[1] 三菱電機株式会社 情報技術総合研究所
{Endo.Takao@cj, Okabe.Kenta@ah, Ishigami.Shotaro@ap, Nakamura.Mitsuki@bc,
Nidaira.Masaya@bk, Otomura.Kotaro@df, Hato.Jumpei@ea}.MitsubishiElectric.co.jp

概要

VLM の工業製品画像に対する認識性能を調査するため，MVTec-AD データセットに含まれる 15 種類の工業製品画像を使い VQA タスクを実施した．実験では，VLM に被写体とその状態に対する質問と回答の選択肢を入力し，VLM の回答の再現率を評価した．その際，参考画像を与える場合と与えない場合の二通りを評価した．被写体に関する質問では，参考画像を与えなくても高い再現率となる製品があることや，参考画像を与えることで再現率が上昇する製品があることを確認した．一方で，参考画像を与えることでかえって再現率が低下する製品も存在した．被写体の状態に関しては，参考画像を与えることで回答の再現率は上昇する傾向があるが，全体的には被写体を回答することに比べ再現率は低いことを確認した．

1 序論

近年，テキストだけではなく，画像や音声といった複数のモダリティに対応した大規模言語モデル (LLM) が登場している．これらはマルチモーダル LLM と呼ばれ，複数の入力を受け取ることでより人間に近い挙動を再現することが期待され，これまでの LLM をより賢くすると考えられている [1]．マルチモーダル LLM の中でも，画像とテキストを同時に処理可能なものは Vision-Language-Model(VLM) とも呼ばれ，商用・オープンソース含め多数のモデルが存在している [2, 3, 4]．

VLM の活用例として，画像を入力とした異常検知が挙げられる．従来の AI モデルによる画像を使った異常検知では，異常の場所や異常スコアを算出するのみであったが，VLM の登場によりテキストによる出力や，ユーザーが知りたい情報をプロンプトで直接 VLM 問うことで，細かい調整なしにユーザーの所望の処理が可能となった．

また，従来の画像を使った異常検知においては，事前に大量の学習データを用意しておき，検知の目的ごとにモデルの学習を行うことが通例だった．この作業には異常を含む膨大な画像データの収集や，それぞれの画像に対するアノテーション作業など，多大なコストが発生する．これに対し，学習済みの VLM を活用することで，このようなコストを伴わずに画像の異常検知が可能となりつつある．

製造業においては，製品画像と学習済みの VLM を用いた低コストな製品検査が期待されている．しかしながら，学習済みの VLM はインターネット上で収集された画像を学習しているため，製造業におけるドメインに特化した画像に対しては正しく認識できず，推論能力が低いことが懸念されている．

[5] は商用やオープンソースを含めたマルチモーダル LLM の工業製品データセットに対する推論性能を Visual Question Answering (VQA) タスクにより評価している．[5] によれば，評価したモデルの中で最高性能を発揮したのは Open-AI 社の GPT-4o[3] であるが，その推論性能は産業分野で求められる水準には達していないことが報告されている．ただし，[5] では VQA タスクの質問と回答の選択肢を VLM で生成しており，これでは質問と回答の選択肢のセットが生成したモデルに依存するため，最終的な性能評価に不確実性が生じる．そこで本研究では，よりシンプルで機械的に作成した質問と回答の選択肢を用いることで，不確実性を少なくした状態で VLM の工業製品画像に対する認識性能を評価する．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

2 データ

このセクションでは，本研究に使用した工業製品画像データについて説明する．

2.1 MVTec AD dataset

本研究で使用する工業製品の画像データは産業分野の画像を使った異常検知手法のベンチマークデータセットの The MVTec anomaly detection dataset (MVTec AD)[6, 7] である．このデータセットは 15 種類の工業製品それぞれについて学習 (Train) 用画像と評価 (Test) 用画像から構成される．学習用画像は正常 (good) カテゴリのみで構成されており，評価用画像は正常カテゴリと欠陥 (defect) カテゴリの画像で構成されている．工業製品ごとの欠陥カテゴリの数や，各カテゴリに含まれる画像の数は異なり，表 1 に各カテゴリに含まれる画像の枚数をまとめている．本研究では，評価用画像のみを使用した．

表 1: MVTec AD dataset に含まれる画像の概要．[6] をもとに改変して掲載．

	Category	# Train (good)	# Test (good)	# Test (defective)	# Defect groups
Textures	Carpet	280	28	89	5
	Grid	264	21	57	5
	Leather	245	32	92	5
	Tile	230	33	84	5
	Wood	247	19	60	5
Objects	Bottle	209	20	63	3
	Cable	224	58	92	8
	Capsule	219	23	109	5
	Hazelnut	391	40	70	4
	Metal Nut	220	22	93	4
	Pill	267	26	141	7
	Screw	320	41	119	5
	Toothbrush	60	12	30	1
	Transistor	213	60	40	4
	Zipper	240	32	119	7
	Total	3629	467	1258	73

3 方法

このセクションでは，VLM の工業製品画像に対する認識性能を評価する方法について説明する．VLM の画像認識性能を測る指標に CLIPScore[8] がある．これは画像に対するキャプション生成を通じて，事前に画像とキャプションのペアを学習した CLIP と呼ばれるモデルによりスコアを算出するものである．しかしながら，CLIP 自体が工業製品画像について十分に学習しているかは不明なため，本研究においては算出されたスコアがどれだけ信頼できるか不明である．

そのため本研究では，VQA と呼ばれる，画像とその画像に対する質問を与えた際に正しい答えを回答するタスクにより，VLM の工業製品画像に対する認識性能を評価する．VLM には Open-AI 社の GPT-4o[3] を使用し，VQA における質問と回答の選択肢には，TIFA[9] のフレームワークを採用した．TIFA のフレームワークについては次のセクションで説明する．

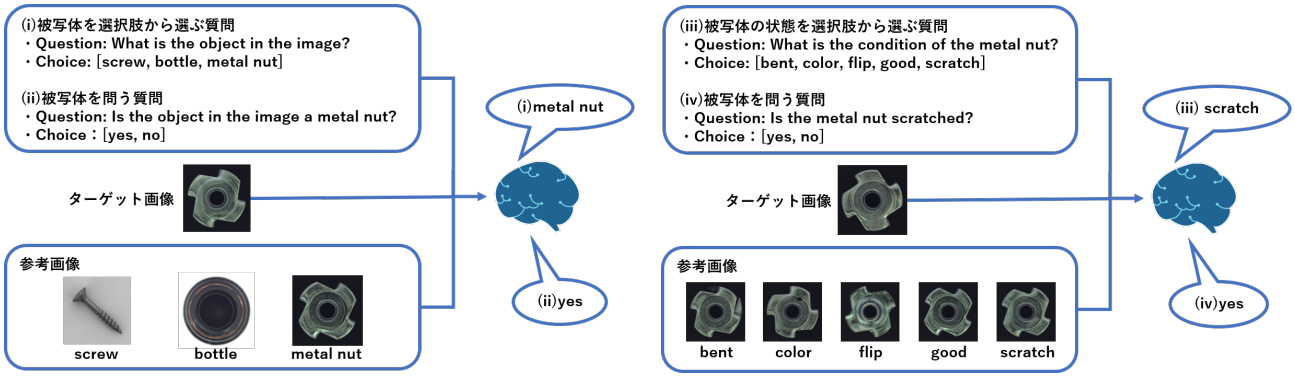
3.1 VQA

VQA は画像の内容について質問を行い，正しい答えを導き出すタスクである．本研究ではその中でも TIFA[9] のフレームワークを用いた．TIFA はもともと画像生成において，意図した画像が生成されているかを評価するフレームワークである．TIFA では，画像生成における入力のプロンプトに含まれる要素に対し質問と回答の選択肢を言語モデルにより生成する．質問には二つのパターンが存在し，一つは画像に写っている要素を選択肢から選ぶものであり，もう一つは画像に要素が写っているか否かを選ぶものである．この質問と回答の選択肢を画像と共に VLM に入力し，出力された VLM の回答と事前に用意した正解を比較して生成画像がプロンプトのとおりに生成されているかを評価する．

TIFA 自体は多数のプロンプトと生成画像に対し，生成された画像が意図通りであるかを自動で評価するフレームワークだが，画像の内容が既知である場合には VLM モデルの認識性能を評価するために使用可能である．すなわち，本研究において使用する MVTec の工業製品画像については，被写体やその状態については既知のため，VLM に画像に写っている被写体やその状態について質問し，正しく回答できるかを調べることで，VLM が工業製品画像についてどの程度正しく認識できているか評価することができる．

図 1 はターゲットの画像を metal nut とした際の VQA の例である．図 1a は被写体に関する VQA であり，被写体を選択肢から回答する質問 (i) と被写体が metal nut か否かを回答する質問 (ii) をターゲット画像とともに VLM に入力している．なお，図 1a の (i) における回答の選択肢は [screw, bottle, metal nut] の三つであるが，実験では MVTec のデータに含ま

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).



(a) 被写体に関する VQA　　**(b)** 被写体の状態に関する VQA

図 1: TIFA フレームワークにおける VQA の例．図 1a では三つの製品を選択肢としているが，実験では全ての製品を選択肢としており，参考画像を与える場合も全ての製品の good 状態の画像を 1 枚ずつ入力している．

れる 15 製品全ての製品名を選択肢として入力した．同様に参考画像を与える場合には 15 種類の製品の good 状態の画像を一枚ずつ入力した．

図 1b は被写体の状態に関する VQA の例である．(iii) では metal nut の状態を選ぶ質問と回答の選択肢を入力しており，回答の選択肢には metal nut の全ての状態を選択肢として入力した．(iv) では metal nut の状態がターゲット画像の状態である scratch か否かを問う質問と回答の選択肢を入力している．被写体の状態について参考画像を与える場合には，その被写体のそれぞれ状態の画像を 1 枚ずつ，状態の名前と併せて入力した．

3.2 評価方法

このセクションでは VLM の回答の評価方法について説明する．今回使用する画像の内容は既知であることから，画像の内容を正しく認識できていたかを測る指標として再現率

$$\text{再現率} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

を評価指標として採用した．ここで TP は被写体やその状態を正しく回答した数であり，FN は被写体やその状態を正しく回答出来なかった数である．

VLM の回答の再現率は，画像の被写体やその状態により変動することが予想される．そのため，被写体の画像全体の再現率とともに，被写体の状態毎に再現率も算出した．

4 結果

このセクションでは，ターゲット画像の被写体に関する質問 (i)，(ii) と被写体の状態に関する質問 (iii)，(iv) の回答を評価した結果を述べる．なお，全ての工業製品に対する再現率の結果は付録 A の図 3，図 4 に掲載している．

4.1 被写体に関する質問

画像の被写体に関する質問は，(i)「被写体を選択肢から回答する質問」と (ii)「被写体が問い合わせたものであるか回答する質問」の二種類ある．図 2a は metal nut と zipper について，それぞれの質問の回答の再現率を参考画像の有無により比較したものである．この図では各被写体のそれぞれの状態と，被写体の画像全体について算出した再現率をレーダーチャートで表示している．metal nut においては，参考画像がない場合に比べ，参考画像を入力した場合の方が再現率が上昇したことが読み取れる．特に (ii) の被写体が metal nut かどうかの質問に対しては参考画像を入力することで再現率が大きく上昇している．metal nut のように参考画像を与えた場合に再現率が上昇した製品は carpet, tile, cable だった．

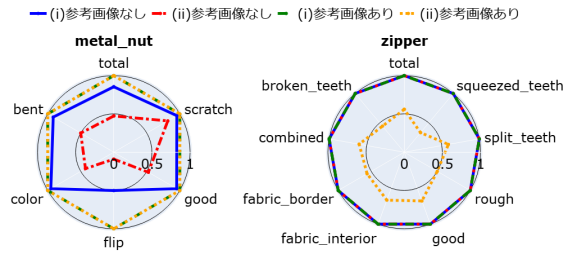
一方 zipper に関しては (ii) の zipper であるかという質問に対しては参考画像を入力した場合，回答の再現率が減少した．参考画像がない場合には，どの状態においても再現率は 1 だったため，予想外の結果である．zipper の他にも，leather, wood, transistor において，同様の結果が確認された．

他の被写体に関しては，参考画像の有無による再現率の変化はあまりなく，かつ参考画像なしの場合でも高い再現率だった．

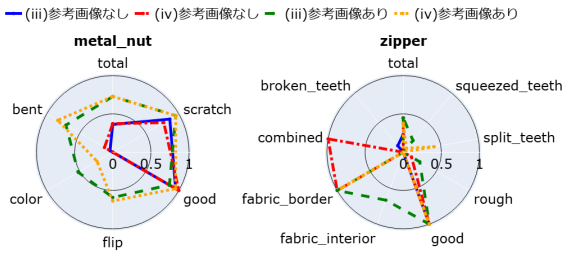
4.2 被写体の状態に関する質問

画像の被写体の状態に関する質問は，(iii)「被写体の状態を選択肢から回答する質問」と (iv)「被写体の状態が問い合わせたものであるか回答する質問」の二種類ある．図 2b は metal nut と zipper につ

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).



(a) 被写体に関する質問の再現率



(b) 被写体の状態に関する質問の再現率

図 2: metal nut と zipper に対する VLM 回答の再現率．左は被写体に関する質問．右は状態に関する質問．

いて，それぞれの質問の回答の再現率を参考画像の有無で比較したものである．metal nut では，参考画像を入力することで各状態の回答の再現率は上昇，または，あまり変わらないかのどちらかだったことが読み取れ，metal nut 全体としては回答の再現率は上昇している．参考画像を入力した場合にそれぞれの状態で回答の再現率が上昇または変化のない製品は grid,leather, bottle, cable, capsule, hazelnut, toothbrush, transistor だった．

zipper に関しては，参考画像を与えた場合には回答の再現率が上がる状態もあれば，逆に大きく下がる状態も存在した．zipper 全体としては参考画像の有無により回答の再現率は若干上昇した．zipper のように，参考画像を入力することで，再現率が下がる状態を含む製品は carpet, tile, wood, pill, screw だった．それぞれの製品全体としては参考画像を入力した場合には再現率が上昇する傾向にあった．

5 考察

セクション 4.1 では，参考画像を与えなかった場合，(i) と (ii) とで回答の再現率が大きく異なる製品が存在した．例えば，図 2a の metal nut においては (i) の質問では回答の再現率が高い状態でも，(ii) の質問では回答の再現率が低下している．同様の現象は carpet でも確認され，こちらは metal nut よりも顕著に低下している．このことから，15 種類の製品の選択肢の中から選ぶ場合には消去法的に正解を選ぶことが出来ているが，本来は別のものとして VLM が認識していた可能性がある．それを確かめるためには，例えば (i) の質問の回答の選択肢に，「その他」という選択肢を設けたり，選択肢の中に VLM が考える正解がない場合には VLM が被写体をどのように認識しているかを回答させ，分析したりすることが挙げられる．これらは今後の課題として取り組む予定である．

また，参考画像を与えた場合には，(ii) の質問で回答の再現率が低下する現象も確認された．原因としては，今回の実験では (i) と条件をそろえるため，(ii) の質問でも 15 種類の画像全てを入力したことが挙げられる．(ii) の質問は被写体が問い合わせたものであるか否かという内容のため，参考画像としては同じ被写体の画像を用いれば十分であるが，多種類の画像を入力したことで VLM が混乱した可能性がある．参考画像を限定することで，VLM から見ても参考画像と同じ被写体かどうかというシンプルなタスクになり，より本来の性能が試されると考えられる．これについても，今後の課題として取り組む予定である．

6 結論

VLM の工業製品画像に対する認識性能を調査するため，TIFA のフレームワークにより VQA タスクを行い VLM の回答の再現率を評価した．その際，参考画像を含める場合と含めない場合とで結果を比較した．被写体を回答する (i),(ii) の質問では，回答の再現率は全体的に高かったが，参考画像が無い状態では (ii) の回答の再現率が (i) に比べ低下したり，あるいは参考画像がある場合の方が回答の再現率が低下する製品が確認された．このことから，VLM は消去法で正解を答えていたが，被写体を別のものとして認識していた可能性が示唆された．被写体の状態を問い合わせる (iii), (iv) の質問では，参考画像を入力することで回答の再現率が低下するものも見られたが，多くの場合で回答の再現率が上昇することを確認した．しかしながら，被写体に対する回答の再現率に比べると，全体的に回答の再現率は低い傾向が存在した．このことから，被写体の状態に対する認識性能はそれほど高くないことが確認された．以上から VLM の工業製品に対する本来の認識性能への懸念は正しいことが示された．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. **arXiv e-prints**, p. arXiv:2306.13549, June 2023.

[2] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2024.

[3] OpenAI. GPT-4o System Card. **arXiv e-prints**, p. arXiv:2410.21276, October 2024.

[4] Gemini Team Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. **arXiv e-prints**, p. arXiv:2403.05530, March 2024.

[5] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. MMAD: The First-Ever Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection. **arXiv e-prints**, p. arXiv:2410.09453, October 2024.

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection, 2019.

[7] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. **International Journal of Computer Vision**, Vol. 129, No. 4, pp. 1038–1059, Apr 2021.

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. **arXiv e-prints**, p. arXiv:2104.08718, April 2021.

[9] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. **arXiv preprint arXiv:2303.11897**, 2023.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　各工業製品に対する再現率

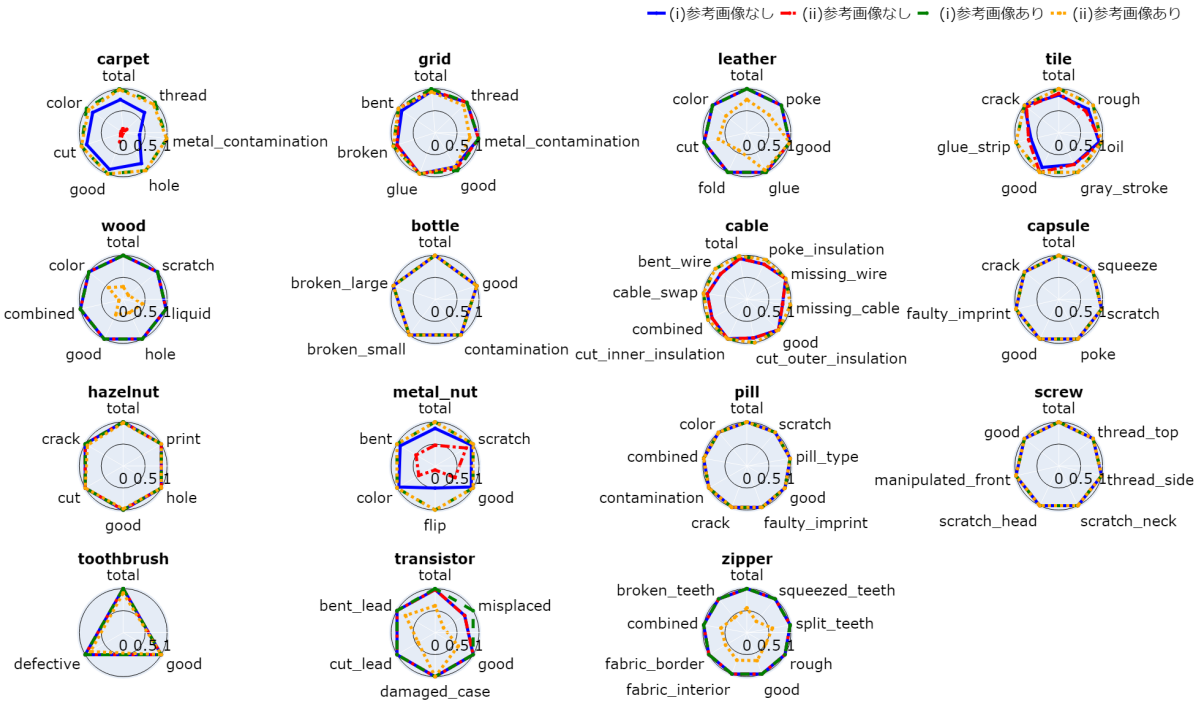


図 3: MVTec の工業製品画像について，被写体に関する質問 (i)(ii) を行った際の参考画像の有無による再現率の比較.

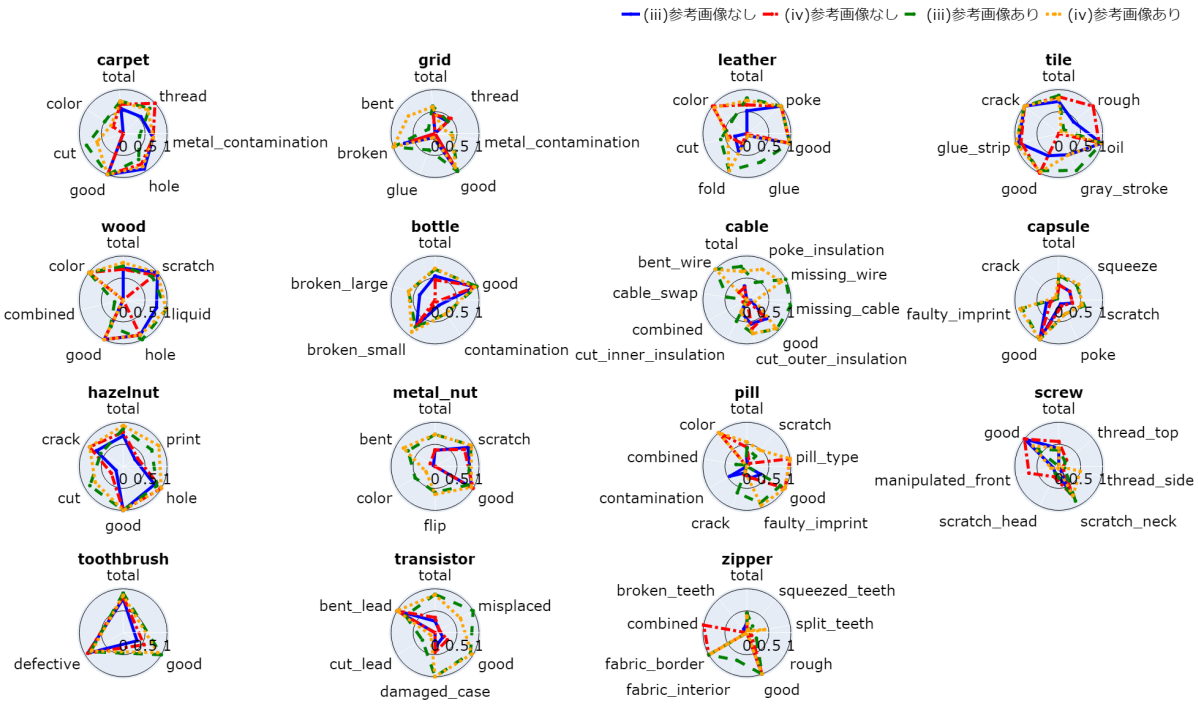


図 4: MVTec の工業製品画像について，被写体に関する質問 (iii)(iv) を行った際の参考画像の有無による再現率の比較.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).