VLM を用いたドメイン特化生成画像の定量評価

岡部 健太[1] 遠藤 隆夫[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
{okabe.kenta@ah, endo.takao@cj, ishigami.shotaro@ap, nakamura.mitsuki@bc,
nidaira.masaya@bk, otomura.kotaro@df, hato.jumpei@ea}.mitsubishielectric.co.jp

概要

視覚言語モデル（VLM）の進歩に伴い，画像品質評価においてVLMを用いることが注目されている．また，製造業では製品の外観検査において異常画像の数が少ないため精度の高い異常検知モデルの開発が難しく，異常画像を生成できるドメイン特化画像生成 AI の需要が高い．そこで本研究では VLM を用いてドメイン特化生成画像の定量評価を行い，その妥当性を検証した．VLM により生成画像を定量評価し，既存評価値や官能評価値との相関を分析した結果，特に形状，色，テクスチャの観点で VLM によるドメイン特化生成画像の定量評価の可能性を示した．また，参照画像や生成画像同士の相対評価を行うことで，精度の高い評価が可能となることを示唆した．

1 はじめに

近年，画像生成 AI 技術が急速に普及し，高品質な画像を生成することが可能となっている．特定の分野や用途に特化したドメイン特化画像生成 AI の研究[1]も進められており，より専門的なニーズに応えることが期待されている．製造業では製品の外観検査タスクにおいて，製品の異常画像は正常画像と比較して非常に入手しにくく，異常検知モデルの精度を高めることが難しいという課題がある．そのため，一般的な人物や風景を生成する汎用的な画像生成 AI ではなく，製品の異常画像を生成するようなドメイン特化の画像生成 AI の需要が高い．

また，テキストと画像を同時に扱うマルチモーダルな視覚言語モデル（VLM: Vision Language Model）も進化している．近年，AI で生成した画像の品質を評価するため，VLM を用いることが注目されている．VLM では従来の評価指標では捉えきれないような画像の細かな品質や文脈に応じた評価ができる可能性がある．また，VLM では定量評価だけでなく定性評価も可能なので，既存の評価指標と比べて個々の画像について詳細な分析がしやすいといった利点もある．

先行研究[2]において VLM を用いることで画像品質を定量評価できる可能性が示されており，本研究では VLM による画像の品質評価がドメイン特化の生成画像に適用できるか検討した．これによりドメイン特化生成画像の評価プロセスの確立および評価精度向上を目指す．

2 既存の生成画像評価方法

画像生成は VAE（Variational Autoencoder）[3]や GAN(Generative Adversarial Network)[4], Diffusion モデル[5]をベースとした様々なモデルが考案されている．それに伴い生成画像の評価方法もまた様々な方法があり，特定の観点における指標を用いた定量評価，人間が視覚的に評価する主観評価，後段の生成画像を用いる分類タスクや検知タスクなど特定のタスクにおける精度評価等がある．現在は特定観点における指標を用いた定量評価が主流であり，生成画像の多様性と品質を評価する IS（Inception Score）[6]や生成画像と実画像の分布の違いを評価する FID（Frechet Inception Distance）[7]，輝度，コントラスト，構造の観点から生成画像と実画像の類似度を比較する SSIM（Structural Similarity）[8]などがある．

近年では VLM の普及に伴い，VLM を用いた画像品質評価について様々な研究がされている．VLM 向けの画像品質評価向けのベンチマークデータセット Q-Bench [9]では，VLM がどれだけ画像に対して正確に品質を回答することができるかといった評価や定量的な画像品質スコアを予測するベンチマークを提案している．VLM による画像品質評価の可能性を調査した研究では，gpt-4v において人間の画像品質の知覚を適切に説明できる可能性が示唆されている[2]．しかし，一般ドメインは検証されているが，VLM で学習されていないような特定ドメインにおける画

像では十分に検証されていない.

本研究では, VLM を用いてドメイン特化生成画像の定量評価を行い，その妥当性を検証することを目的とする.

3 実験

3.1 実験概要

実験は VLM によるドメイン特化生成画像における定量評価の妥当性を検証することを目的とする. 生成画像を VLM で複数の評価観点でスコア化し，スコアと既存の評価指標や官能評価値との相関係数を算出することで，VLM 評価の妥当性を検証した.

VLM で定量評価する生成画像については, 異常検知タスク向けのデータセット MVTec [10]の Metal nut を学習して生成した画像を用いた. 具体的には Stable Diffusion XL[11]に対して正常（good），異常(scratch)画像を用いて LoRA[12]により傷の位置の異なるモデルを 2 つ学習し，Multi LoRA[13]により画像を生成した. 生成画像は正常画像を 400 枚，異常画像は Multi LoRA の 3 パターン（merge, switch, composite）をそれぞれ 400 枚ずつ，トータルで生成画像 1600 枚を評価した.

先行研究では OpenAI のモデルである gpt-4v が最も高い精度を示しているとされている[2]. 本研究では，その後に登場したモデルである gpt-4o を定量評価させる VLM として選定した.

gpt-4o では一般的なデータは学習されている一方で，ドメイン特化のデータは学習されていない可能性が高い. 事前情報なしでドメイン特化画像を評価させるのは難しいため，本実験ではオリジナル画像を参照画像として生成画像と同時に VLM に入力した. 本研究では図 1 のように参照画像 1 枚，生成画像 1 枚を VLM へ入力して 1 枚の生成画像を評価する単一生成画像評価と，参照画像 1 枚，生成画像 4 枚を VLM へ入力して生成画像 4 枚を同時に評価する複数生成画像評価の 2 つの方法で評価した.

3.2 単一生成画像評価法

3.2.1 実験条件

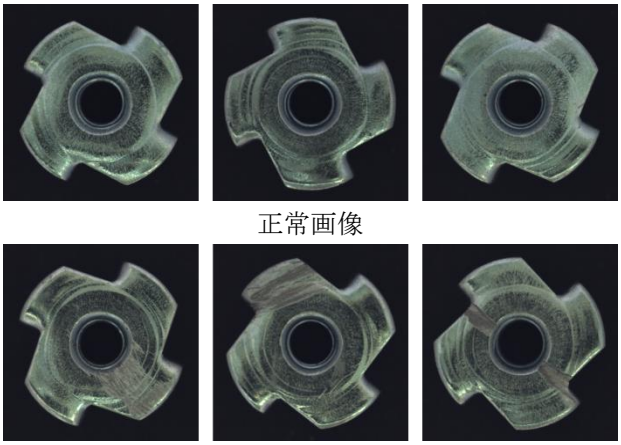
単一生成画像評価法では参照画像 1 枚と生成画像 1 枚を VLM へ入力し，4 つの観点で生成画像を 0-100（100 を参照画像の点数）のレンジで参照画像にどれだけ近いか品質を評価させた. 評価観点は形状，色，テクスチャ，傷（異常画像のみ）とした. 参照画像はファインチューニングに用いたオリジナル画像から図 2 の正常画像，異常画像をそれぞれ 3 枚ずつ選んだ. 1 枚の生成画像につき 3 枚の参照画像におけるそれぞれのスコアを算出し，その平均値を用いた. 3 枚の参照画像はナットの色，角度，テクスチャ，傷（異常画像のみ）の観点から幅広く選定した.

表 1　評価画像枚数

画像種類		枚数【枚】
good		400
scratch	merge	400
	switch	400
	composite	400

(i)単一生成画像評価法　　(ii)複数生成画像評価

図 1 VLM への画像入力



正常画像

異常画像

図 2 参照画像

3.2.2 VLM による定量評価妥当性検証

VLM による評価の妥当性を検証するために, 評価観点に関連する指標を算出し, 相関係数を算出した. 形状は Hu モーメント[14], テクスチャは GLCM[15], 色はヒストグラム，傷は官能評価値を用いて，相関係数を算出した.

(1)Hu モーメント：画像の形状特徴を表すための指標であり回転，スケーリング，平行移動に対して影

響を受けない指標である．1 枚の生成画像に対し，すべてのオリジナル画像と Hu モーメントの差分値を算出し，その差分値の平均値と VLM による形状のスコアの相関係数を導出した．

(2)ヒストグラム：画像内の色分布を表現するために画像の各ピクセルの色の出現頻度をカウントする手法である．本検証ではヒストグラムの重なり具合を比較するインターセクション法[16]で類似度を算出する．1 枚の生成画像に対し，すべてのオリジナル画像との間で類似度を算出し，その類似度の平均値と VLM による色のスコアの相関係数を導出した．

(3)GLCM：画像のテクスチャ特徴を抽出するための指標である．GSML は回転要素に影響を受けること，またオリジナル画像，生成画像ともにナットの方向が様々であるため，本検証ではそれぞれの画像を 4 方向（0°, 45°, 90°, 135°）ずつの計 16 通りの組み合わせで GLCM からユークリッド距離を算出し，そこから類似度を導出する．1 枚の生成画像に対し，すべてのオリジナル画像との間で類似度を算出し，その類似度の平均値と VLM によるテクスチャのスコアの相関係数を導出した．

(4)官能評価：ナットの傷の評価に適した既存指標がないため，傷は官能評価値を用いた．異常生成画像 400 枚のうち 100 枚をランダム抽出し，2 名の評価者が表 2 の評価基準に沿って生成画像がオリジナル画像とどれだけ似ているかを 1-5 で評価した．本検証では 2 名の評価者による傷の官能評価の平均値と VLM による傷のスコアの相関係数を導出した．

上記の方法により算出した相関係数を表 3 に示す．なお，表 3 では VLM で正しく評価できていれば正の相関になるように形状とテクスチャの相関係数の正負を反転させた．また，merge において色に関するスコアがすべて 95 であったため，相関係数は算出不可である．表 3 より色は composite で，テクスチャは composite と switch で相関係数が 0.4 以上の正の相関があり，VLM により正しく評価できている．逆に傷については switch で-0.34 と弱い負の相関があり，正しく評価できていない．ただし，単体画像評価においては VLM によるスコアのレンジが狭く，特に異常画像（composite, switch, merge）の傷や，merge の色，正常画像のテクスチャは VLM によるスコアのレンジが狭いため，正確な評価ができずに相関係数が小さくなった可能性がある．そこで複数生成画像評価法で相対評価にすることにより正確な評価になると考えられる．

表 2　官能評価における評価基準

スコア	評価基準
5	オリジナル画像と非常に似ており，品質は非常に高い．
4	オリジナル画像とかなり似ており，品質は良い．
3	オリジナル画像とある程度似ており，品質は普通．
2	オリジナル画像と少し似ているが，品質は低い．
1	オリジナル画像と全く似ておらず，品質は非常に低い．

表 3　単一生成画像評価における相関係数

		評価観点			
		形状	色	テクスチャ	傷
good		0.009	0.058	0.039	対象外
scratch	composite	0.156	0.437	0.443	-0.113
	switch	0.022	0.195	0.630	-0.340
	merge	0.002	-	-0.199	-0.094

3.3 複数生成画像評価法

3.3.1 実験条件

複数生成画像評価法では参照画像 1 枚と生成画像 4 枚を VLM へ入力し，VLM に 0-100（100 を参照画像の点数）のスコアで参照画像にどれだけ近いか生成画像 4 枚の品質を評価させた．評価観点や参照画像は単一生成画像評価法と同条件とした．生成画像 4 枚は同条件 400 枚からランダムの組み合わせとし，3 枚の参照画像それぞれ別の組み合わせで評価した．

3.3.2 VLM による定量評価妥当性検証

単一生成画像評価法では VLM スコアと既存評価指標や官能評価値の相関係数を求めたが，複数生成画像評価法では相対評価であるため，スコアではなく相対順位を用いた．VLM へ同時に入力した 4 枚の生成画像の中で VLM のスコアが最も高い生成画像のランクを 1，スコアが最も低い生成画像のランクを 4 となるようにランクを割り当て，参照画像を変化させた 3 回分の平均ランクを算出した．平均ランクと既存評価指標や官能評価値との相関係数を算出し，VLM による定量評価の妥当性を検証した．

表 4　複数生成画像評価における相関係数

		評価観点			
		形状	色	テクスチャ	傷
good		0.166	-0.072	0.208	対象外
scratch	composite	0.234	0.583	0.734	-0.255
	switch	0.397	0.165	0.532	-0.217
	merge	0.127	0.143	-0.231	0.094

単一生成画像評価法と同様に形状はHuモーメント，テクスチャはGLCM，色はヒストグラム，傷は官能評価値を用いて相関係数を算出した．

複数生成画像評価法における相関係数を表 4 に示す．なお，表 4 では VLM で正しく評価できていれば正の相関になるように色と傷の相関係数の正負を反転させた．表 4 から形状においては composite で 0.234，switch で 0.397 と弱い正の相関となり，色は composite で 0.583, テクスチャは composite と switch でそれぞれ 0.5 以上と正の相関が得られた．一方，傷については composite と switch でそれぞれ-0.255, -0.217 と弱い負の相関となった. 条件によるが, 形状，色，テクスチャについては VLM により概ね正しく評価できているといえる．

4 考察

単体生成画像評価法と比べると，複数生成画像評価法では多くの条件において相関係数が大きくなった．特に単体生成画像評価法では相関が低かった形状では相関係数が高くなり，評価が改善された．今回は参照画像を 100 点として VLM へ提示したが，その他の点数の画像は提示していない．100 点未満の点数の基準は VLM に委ねられるため，単体生成画像評価では多少の違いでは VLM のスコアに差が表れにくく，複数生成画像評価法では他の生成画像との相対評価できるようになり，細かい違いを評価しやすくなったと考えられる．単体画像評価法でも詳細な違いを反映させるためには，参照画像と同様に，例えば 80 点の基準画像とスコアを VLM へ同時に入力することで対応できると考えられる．

評価観点別では条件によるが形状，色，テクスチャは正の相関，傷は負の相関の傾向となった．傷において負の相関が発生した要因について考察する．官能評価では高品質と評価した傷が目立つ生成画像は VLM では低いスコアとなり，官能評価では低品質と評価した傷が目立たない生成画像を VLM では高いスコアをつける傾向があった（付録 A 参照）．VLM にスコアの理由を出力した結果, 参照画像と比較した傷の視認性を 1 つの評価基準としており，参照画像より傷が目立たない画像を高品質，参照画像より傷が目立つ画像を低品質と判断していることを確認した．傷があると品質が悪いといった一般的な品質の定義が VLM の傷のスコアにも反映され，結果的に官能評価と VLM のスコアが負の相関になったと考えられる．今後，VLM で傷を評価する場合は評価基準をプロンプトに明確に入力必要がある．

また，条件別では composite と switch は相関係数が高く，good と merge では相関係数が低い傾向であった．composite と switch は背景色がグレーや赤色で，ナットの形状が明らかに異なる画像が多く生成された．一方，good や merge は good の形状を除いて，明らかにオリジナル画像と異なる生成画像は少なかった．そのため，オリジナル画像との違いが大きい画像に関しては VLM で正しく評価でき，違いが小さい画像については正確に評価できていない可能性がある．今後は他の指標を用いた画像分析を進めることで条件によって傾向が異なった要因分析を進めるとともに，どこまで微細な変化を VLM で評価できるか検証を進める．

5 おわりに

本研究では VLM を用いてドメイン特化生成画像の定量評価を行い, VLM による評価の妥当性を検証した．

単一生成画像評価法では色やテクスチャの評価観点で一部の条件において正の相関が得られたが，VLM のスコアのレンジが小さく, 正確に評価できていない可能性があった．そこで複数生成画像評価法でも同様に色，テクスチャに加えて形状も正の相関が確認された．以上から，形状，色，テクスチャの観点において VLM によるドメイン特化の生成画像の定量評価の可能性を示した．また，ドメイン特化の画像では VML に評価させたい画像だけでなく，参照画像や生成画像同士の相対評価をさせることで，より精度の高い評価になることを示唆した．

今後は，条件によって異なる傾向が見られた要因を分析するとともに，参照画像や基準点数の画像を複数提示することや明確な評価基準をプロンプトで明示することでより正確な評価を目指し，どの程度の違いまで VLM で評価できるか検証を進める．

参考文献

1. Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, Chengjie Wang, AnomalyDiffusion: Few-Shot Anomaly Image Generation with Diffusion Model, Proceedings of the AAAI Conference on Artificial Intelligence, vol.38(8), pp.8526-8534, 2024

2. Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang, A Comprehensive Study of Multimodal Large Language Models for Image Quality Assessment, Lecture, Notes in Computer Science, vol.15132, pp143-160, 2024

3. Diederik P Kingma, Max Welling, Auto-Encoding Variational Bayes, arXiv:1312.6114, 2013

4. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative Adversarial Nets, NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol.2, pp.2672-2680, 2014

5. Jonathan Ho, Ajay Jain, Pieter Abbeel, Denoising Diffusion Probabilistic Models, arXiv:2006.11239, 2020

6.Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, Improved Techniques for Training GANs, NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp.2234-2242, 2016

7.Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, Advances in Neural Information Processing Systems 30, pp.6629-6640, 2017

8. Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE transactions on image processing, vol.13(4), pp.600-612, 2004

9. Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, Weisi Lin, Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision, ArXiv:2309.14181, 2023

10. Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger and Carsten Steger, MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection, International Journal of Computer Vision, vol.129, pp.1038-1059, 2021.

11.Stable Diffusion, stabilityai/stable-diffusion-xl-base-1.0, （2024-10 閲覧）, https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0/blob/main/README.md.

12. Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, LoRA: Low-Rank Adaptation of Large Language Models, arXiv:2106.09685, 2021

13. Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, Weizhu Chen, Multi-LoRA Composition for Image Generation, arXiv:2402.16843, 2024

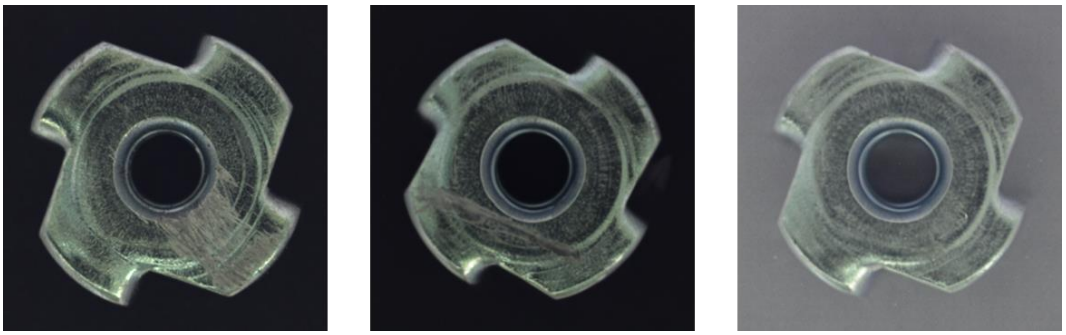
14. Ming-Kuei Hu, Visual Pattern Recognition by Moment Invariants. IRE Transactions on Information Theory, vol.8(2), pp.179-187.

15. Robert M. Haralick; K. Shanmugam, Textural Features for Image Classification, IEEE Transactions on Systems, Man, and Cybernetics, vol.SMC-3(6), pp.610-621, 2007

16. OpenCV 公式ドキュメント Histograms, （2024-11 閲覧）, https://docs.opencv.org/4.9.0/d6/dc7/group__imgproc__hist.html

A VLM による傷の定性評価分析

評価観点別では傷のみ VLM のスコアと官能評価のスコアが負の相関の傾向となったため，VLM にスコアを算出した理由を出力し，負の相関となった理由を分析した．VLM における参照画像を図 3 の(A)とした場合，(B)のような傷が目立つ生成画像では VLM では低スコア，官能評価では高スコア，(C)のような傷が目立たない生成画像では VLM では高スコア，官能評価では低スコアをつける傾向があった．この 2 枚の生成画像に VLM にスコアをつけた理由を出力した結果を表 5 に示す．参照画像と比較した傷の視認性を 1 つの評価基準としており，参照画像より傷が目立たない画像を高品質，参照画像より傷が目立つ画像を低品質と判断している傾向があった． 傷があると品質が悪いといった一般的な品質の定義が VLM の傷のスコアにも反映されている可能性がある．



(A)参照画像　(B) VLM：低，官能評価：高　(C) VLM：高，官能評価：低

図 3 参照画像

表 5　傷におけるスコア算出理由

画像	VLM による 傷のスコア	官能評価 平均スコア	VLM によるスコア算出理由
(B)	85	3.5	- The generated image shows a visible scratch across the surface of the nut, which is more pronounced than in the reference image. - The scratch affects the overall visual quality, reducing the score compared to the reference image, which is free of such imperfections. - The rest of the surface appears relatively smooth, but the presence of the scratch is significant enough to impact the score.
(C)	90	1	- The generated image shows fewer visible scratches compared to the reference image. - The surface appears smoother, indicating a higher quality in terms of scratch visibility. - Minor imperfections may still be present, but they are less pronounced than in the reference."