# 産業用画像の異常検知に基づく説明文生成における学習方法の検討

FU YUXUAN[1]　　小林 一郎[1]
[1]お茶の水女子大学大学院 人間文化創成科学研究科
{fu.yuxuan,koba}@is.ocha.ac.jp

## 概要

産業画像理解における異常検知は，これまで主に構造的欠陥や論理的不整合といった異常を画像レベルで解析する手法に焦点が当てられてきた．近年，Vision-Language Models（VLMs）の発展により，言語的推論を組み込むことで，より包括的な異常検知が可能となっている．本研究では，低レベルな視覚特徴に基づく処理と高レベルな言語による推論を組合せた異常検知を実現することを目的として，細粒度なアノテーションを付与した質問応答（QA）拡張データセットに基づく，画像とテキストを統合したマルチモーダル枠組みを提案する．さらに，大規模 VLM を用いた Chain-of-Image（CoI）推論において，異なるプロンプト設計による推論性能を評価し，学習型のマルチモーダル手法と比較を行う．実験の結果，ファインチューニングされたモデルは高い検出精度を達成する一方で，生成される説明の多様性が低下する傾向が確認された．一方，学習不要手法においても，学習あり手法と比較して大きな性能差は見られず一定の競争力を維持できるものの，画像微細な異常の識別には依然として課題が残ることが明らかとなった．

## 1　はじめに

産業画像理解における異常検知は，従来，画像解析に基づく手法が主流であり，特にセグメンテーションに基づく手法は，破損や汚れなどの構造的欠陥の検出において高い性能を示してきた [1]．近年では，製品ラベルと内容物の不一致といった論理的異常を対象とする研究も進展している [2]．一方で，Vision-Language Model（VLM）の発展により，異常検知を視覚情報のみに基づく枠組みに限定することの限界が明らかになりつつある．構造的異常に加え，潜在的な論理的整合性の逸脱を適切に捉えるためには，言語的推論の導入が不可欠である．本研究では，視覚情報と言語情報を統合する学習ベースのマルチモーダル異常検知枠組みを提案し，QA 拡張データセットを用いて VLM を追加学習することで，異常画像および正常画像の特徴を学習する．また，学習不要手法に加え，大規模 VLM を用いた 1-shot の Chain-of-Image（CoI）推論についても検討し，両方法をに比較することで，構造的異常および論理的異常検出におけるそれぞれの有効性と限界を明らかにする．

## 2　関連研究

### 2.1　視覚的異常検知

異常サンプルが十分に得られないという制約から，従来の視覚的異常検知手法では，教師なし学習や自己教師あり学習に基づく枠組みが主に採用されてきた [3]．特に，構造的異常の検出に関しては，これまでに多くの研究成果が報告されている．例えば，SimpleNet [4] は半教師あり異常検知手法の一つであり，学習段階では正常サンプルのみを用い，特徴空間にガウス雑音を加えることで擬似的な異常特徴を生成し，二値分類器によって異常の検出および局所化を行う．また，Kim ら [5] はセグメンテーションモデルを用いて正常画像と異常画像のクラスヒストグラムを比較する手法を提案し，MVTec LOCO データセット [6] において高い検出性能を示している．

### 2.2　マルチモーダル異常検知

CLIP [7] の登場以降，視覚と言語を統合的に扱うモデルが発展し，画像理解はマルチモーダル推論を伴う高度なタスクへと拡張されてきた．さらに，GPT-4 [8] をはじめとする大規模言語モデルの成功により，Qwen [9] や LLaVA [10] などのマ

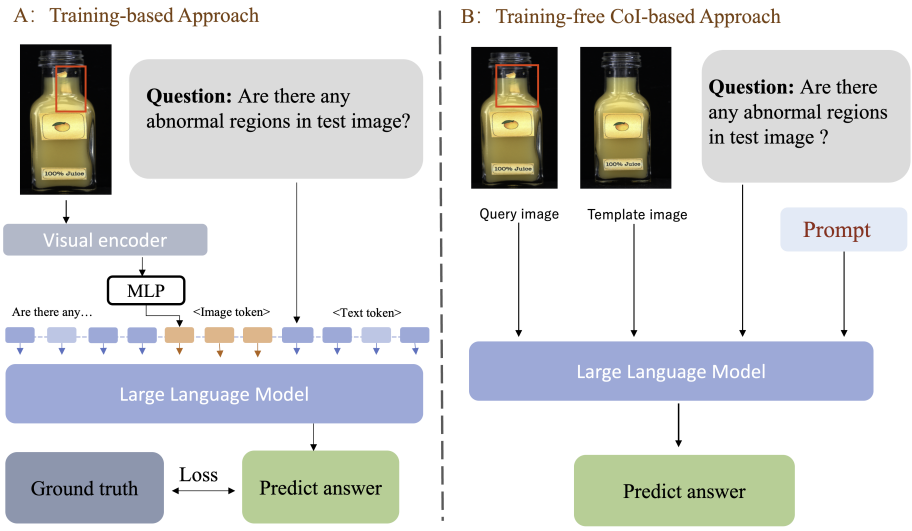


**図 1**　産業画像における異常検知と説明文生成の枠組み

ルチモーダル大規模モデルも，画像理解や Visual Question Answering 分野で顕著な進展を示している．これらの技術は異常検知にも応用されており，Zhang ら [11] は視覚言語モデルを用いた学習不要な異常検知を提案し，高い汎化性を示した．また，AnomalyGPT [12] は異常画像とテキストを用いた学習により，異常検知と位置特定を実現している．一方で，これらの手法は最終出力に焦点を当てることが多く，推論過程の明示的な表現や解釈性には課題が残る．この問題に関連して，Jiang ら [13] は，Wei ら [14] によって提案された Chain-of-Thought の考え方を用い，産業異常検知に特化したベンチマークを構築している．本研究では，これらの先行研究の知見に基づき，Wei らの手法と類似した Chain-of-Image（CoI）推論手法を提案した Meng ら [15] のアプローチを採用し，異常検知への適用可能性を検討する．

# 3　提案手法

本研究では，構築した QA 拡張データセットに基づき，産業画像異常検知のための 2 種類の補完的な方法を提案する．図 1 に示すように，全体の枠組みは以下の 2 つの部分（Part A と Part B）から構成される．**Part A**(学習ベース手法): この方法では，QA 拡張 データセットを用いた教師あり学習を行う．学習には正常サンプルに加え，一部の異常サンプルも使用し，その後，異常検知および推論に適用する．本手法は，学習を行うことでどれだけ性能向上が得られるかを検証することを目的としている．**Part B**(学習不要な方法): 既存のマルチモーダル大規模モデルが一般的に複数画像の比較推論をサポートしている点，そして異常検知が本質的に正常・異常の比較に基づいている点を踏まえ，本研究では学習不要の手法を使用した．本手法ではモデルの追加学習を行わず，QA アノ テーションに基づく 1-shot 推論を直接実行することで，現在のマルチモーダルモデルが持つ異常検知能力を評価する．この 2 つの方法を併用することで，同一のマルチモーダル環境下において，学習ベース手法と学習不要手法が産業異常検知に与える影響を比較することが可能となる．

## 3.1　データセット構築

既存の産業異常検知データセット（例：MVTec-LOCO [6]）は画像レベルの情報に限定されており，精細なアノテーションを欠くため，マルチモーダルモデルにおける視覚と言語の整合的な学習が困難である．そこで本研究では，各画像に対して 2 種類の質問応答（QA）を付与した．1 つは高品質な意味情報を確保するために手作業で作成したテキストであり，もう 1 つは GPT により自動生成したテキストに対して人手で検証を行ったものである．使用したデータセットは 5 種類の製品カテゴリから構成されており，異常タイプとして構造異常および論理異常の 2 種類を含む．また，学習用画像 1,929 件，テスト用画像 1,568 件から構成されている．本研究において生成した QA テキストデータの総数は 6,680 件である．

## 3.2　学習ベース手法

学習ベースの方法では，産業画像と対応する質問を入力とし，LLaVA-1.5 をベースラインモデルとして用いる．画像は視覚エンコーダによって特徴抽出され，得られた画像トークンは軽量な MLP により

**P1:**
Step 1: Identify all objects in the template image and describe their expected appearance.
Step 2: Describe any visible differences between the test image and the template image.
Step 3: Judge whether those differences are defect or normal.

**P2:**
Step 1: The template image follows certain visual rules or regular patterns. **Observe the template carefully and summarize these rules or expected visual characteristics.**
Step 2: Based on these rules, describe the differences between the template image and the test image, **focusing on the left, right, and central regions.**
Step 3: Compare these two images and identify any inconsistencies or anomalies found.
Step 4: Determine whether the test image contains any defects. If yes, describe the specific reason or evidence for your judgment.

**P3:**
Step 1: The template image follows certain visual rules or regular patterns. Observe the template carefully and summarize these rules or expected visual characteristics.
Step 2: **Based on these rules**, describe the differences between the template image and the test image, focusing on the left, right, and central regions.
Step 3: **From the differences identified in Step 2**, determine which ones violate the rules in Step 1 and list them as inconsistencies or anomalies. Briefly explain why each one is anomalous.
Step 4: **Based on the anomalies listed in Step 3,** make a final judgment on whether the test image contains any defects. If it does, summarize the main anomaly and its location as the specific reason or evidence for your judgment.

**図 2**　プロンプト設計の例

**表 1**　論理異常および構造異常に対する各モデルの精度

| Method | Logical-Acc. | Struct-Acc. | Average |
|---|---|---|---|
| Llava1.5-7B | 54.3 | 48.1 | 51.2 |
| Llava1.5-7B-CoI | 52.1 | 50.7 | 51.4 |
| Llava1.5-7B-Training | 60.2 | 52.5 | 56.3 |
| InternVL-8b | 64.6 | 60.4 | 62.5 |
| InternVL-8b-CoI | 63.1 | 59.7 | 61.4 |
| InternVL-8b-Training | 69.6 | 60.3 | 64.9 |
| GPT-4o | 69.9 | 70.1 | 70.0 |
| GPT-4o-CoI | **73.2** | **72.5** | **72.8** |

言語モデルの入力空間へ射影される．これらの画像トークンを QA プロンプトのテキストトークンと結合し，マルチモーダル入力として言語モデルに与えることで回答を生成する．学習時には，生成結果と正答とのクロスエントロピー損失を用いてパラメータを最適化し，視覚特徴と異常に関する意味表現の対応関係を学習する．これにより，正常および異常に関する視覚と言語の特徴表現を獲得し，産業画像異常検知における精度と解釈性の向上を図る．

### 3.3　学習不要な手法

学習不要な手法では，モデルのパラメータ更新を行わず，マルチモーダル大規模言語モデルの推論能力を直接利用して異常検知を実施する．具体的には，Chain-of-Image（CoI）推論を採用し，モデルが段階的な推論過程を通じてテスト画像と正常参照画像を明示的に比較することで異常を判断する．本研究では，学習を伴わない推論設定における性能の違いを詳細に分析するため，複数種類のプロンプトを設計し，それぞれを用いた比較実験を行った．使用したプロンプトの構成および指示内容の違いについては図 2 に示す．本研究で構築したデータセットはカテゴリ数が比較的少ないため，評価時には各テスト画像に対して同一カテゴリの正常画像を参照画像として組み合わせる．最終的に，モデルはテスト画像を正常または異常として総合的に判定し，異常と判断した場合にはその説明を付加して出力する．この学習不要な手法により，モデルは追加学習を行うことなく，複数画像間の意味的整合性を直接比較する CoI 推論を活用できる．その結果，異常検知における判断根拠の解釈性を向上させるとともに，学習不要環境におけるモデルの推論能力を明らかにすることができる．

## 4　実験

### 4.1　実験設定

学習ベースの実験では，部分的に教師ありの設定を採用した．MVTec-LOCO の訓練データには正常サンプルのみが含まれるため，それらをすべて学習に使用し，テストセット中の異常画像の 10% をランダムに抽出して学習データに追加した．残りの異常画像と正常画像から新たなテストセットを構成することで，学習段階では正常分布を十分に利用しつつ，テスト段階では未観測異常に対する検出性能を評価できる設定とした．このデータ分割に基づき，LLaVA-1.5-7B および InternVL-8B に対して LoRA によるファインチューニングを行い，エポック数を 30，学習率を 1e-4 に設定した．学習不要なの実験では，学習ベースと同一のテストセットを用い，CoI プロンプトの有無による性能差を比較するとともに，GPT-4o を 1-shot ベースラインとして評価に加えた．すべての学習および推論は NVIDIA RTX 6000 Ada（48GB）GPU 上で実施し，PyTorch 2.4，CUDA 12.1，DeepSpeed を用いた環境で計算を行った．

### 4.2　評価指標

まず，各モデルの異常検知性能を評価する．本データセットには，論理異常および構造異常の 2 種類の異常が含まれているため，表 1 に示すように，異常カテゴリごとに検出精度を個別に報告し，モデ

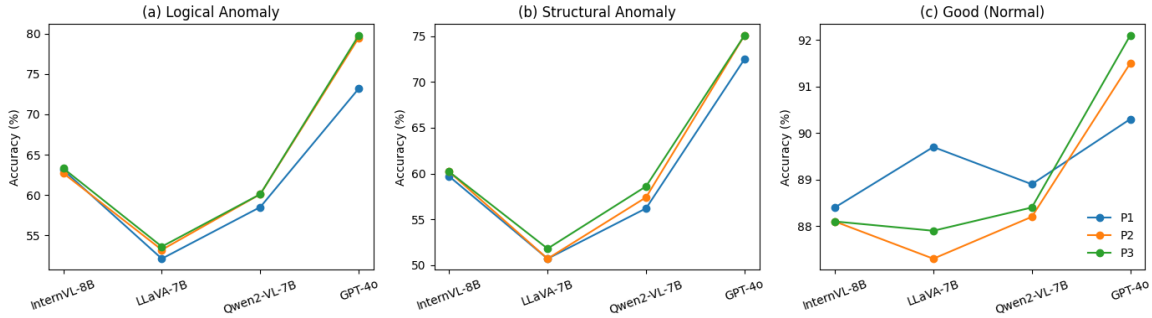


**図 3**　プロンプト変更による検出結果

**表 2**　BERTScore に基づく説明品質の評価結果

| Method | Bert-score |
|---|---|
| Llava1.5-7B | 69.2 |
| Llava1.5-7B-CoI | 70.3 |
| Llava1.5-7B-Training | 83.1 |
| Internvl-8b | 70.3 |
| Internvl-8b-CoI | 75.2 |
| Internvl-8b-Training | 80.5 |
| GPT-4o | 83.2 |
| GPT-4o-CoI | **90.2** |

ルが異なる誤りタイプを適切に識別できているかを検証する．さらに，学習不要な推論手法においては，複数回のプロンプト設計を行い，その影響を比較分析するとともに，既存モデルに加えて Qwen-VL のモデルを導入し，推論性能の違いについても評価を行う．また，学習不要な手法については，正常サンプルに対する判定性能についてもあわせて検証する．具体的には，正常画像を異常と誤って判定していないかに着目し，過検出（false positive）の傾向を分析する．

次に，説明品質の評価を行う．正しく異常を検知したサンプルに対して，生成されたテキストが異常内容を適切に説明できているかを評価する．自動評価指標として BERTScore [16] を採用し，生成文と参照説明文との意味的類似度を算出することで，モデル生成文に正確かつ妥当な異常記述が含まれているかを測定する．

### 4.3　結果と考察

**異常検知性能:** 表 1 に，論理異常および構造異常に対する検出精度を示す．LoRA によりファインチューニングした学習ベースモデル（LLaVA-1.5-7B, InternVL-8B）は，学習段階において正常・異常サンプルを用いた追加学習を行っているため，学習不要な手法と比較して高い検出精度を示した．一方，学習不要な手法においても，一定の性能は確認されたが，学習ベース手法ほどの高い精度には至らなかった．また，GPT-4o は全体を通して最も高い検出精度を達成した．図 3 に示す実験結果から，プロンプト設計は論理的異常の検出性能に影響を与える一方で，構造的異常の検出性能への影響は比較的限定的であることが確認できる．さらに，いずれのプロンプト設定においても，正常（Good）画像の検出精度は一貫して高い水準を維持しており，本研究で採用したプロンプト設計が過検出（false positive）の増加を招いていないことが確認された．

**異常説明の品質:** 表 2 に BERTScore の結果を示す．学習ベースモデルは参照説明との整合性が高く，学習不要な方法より高いスコアを示したが，説明の多様性 はやや低下する傾向が見られた．また，CoI プロンプト を用いることで，すべてのモデルで BERTScore が向上し，推論の安定性や詳細な説明生成に寄与した．

## 5　おわりに

本研究では，QA を付与した異常検知データセットを構築し，部分的に教師あり学習に基づく手法と，CoI プロンプトを用いた学習不要手法という 2 種類の異常検知方法を検討した．論理異常および構造異常の両タスクに対する実験結果から，ファインチューニングしたモデルは高い検出精度を達成する一方で，生成される出力の多様性が低下する傾向が確認された．一方，学習不要の方法は微細な異常の検出には課題を残すものの，依然として競争力のある性能を示すことが明らかとなった．

今後の課題としては，より多様なデータセットの拡張が挙げられる．また，複数の評価指標を導入することで，生成説明の品質を評価するつもりである．

## 謝辞

本研究は科研費（23K28143）に一部支援を頂きました。ここに深謝いたします。

## 参考文献

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 4183–4192, 2020.

[2] Yiheng Zhang, Yunkang Cao, Xiaohao Xu, and Weiming Shen. Logicode: an llm-driven framework for logical anomaly detection. **arXiv preprint arXiv:2406.04687**, 2024.

[3] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In **International conference on pattern recognition**, pp. 475–489. Springer, 2021.

[4] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 20402–20411, June 2023.

[5] Soopil Kim, et al. Few shot part segmentation reveals compositional logic for industrial anomaly detection. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, pp. 8591–8599, 2024.

[6] Michael Bergmann, David Batzner, Klaus Bischof, Markus Fauser, David Sattlegger, and Carsten Steger. The mvtec loco anomaly detection dataset: Towards visually complex anomaly detection. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, pp. 1697–1706, 2022.

[7] Alec Radford, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pp. 8748–8763. PMLR, 2021.

[8] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. **View in Article**, Vol. 2, No. 5, 2023.

[9] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. **arXiv preprint arXiv:2409.12191**, 2024.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **Advances in neural information processing systems**, Vol. 36, , 2024.

[11] Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In **Proceedings of the Computer Vision and Pattern Recognition Conference**, pp. 15204–15213, 2025.

[12] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, pp. 1932–1940, 2024.

[13] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. **arXiv preprint arXiv:2410.09453**, 2024.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, Vol. 35, pp. 24824–24837, 2022.

[15] Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. Chain of images for intuitively reasoning. **arXiv preprint arXiv:2311.09241**, 2023.

[16] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. **arXiv preprint arXiv:1904.09675**, 2019.