構造化知識蒸留と Visual Attention 最適化による ドメイン特化型 VLM の構築

柴田 高志 [1] 田村 颯樹 [1] 月江 惇元 [1] 宮﨑 淳 [1] 髙野 奨太 [1] 中山 和子 [1]
[1]NTT 東日本株式会社
{takashi.shibata.zy, satsuki.tamura.px, atsumasa.tsukie.cw,
jun.miyazaki, takano.s, kazuko.nakayama} @east.ntt.co.jp

概要

近年，VLM の発展により画像とテキストを統合的に理解する AI が実用化されているが，通信設備保守のような特定ドメインでは専門知識やドメイン特化の評価基準への対応が課題である．ドメイン特化型 VLM の開発においては，その開発や評価手法は十分に確立されていないため，本研究では，通信設備保守業務を対象に，構造化知識蒸留と Visual Attention 最適化を統合した VLM 構築手法を提案する．データ面では業務知識を付与した合成データ生成により専門知識を注入し，モデル面では Soft Registers による Attention 制御により視覚的根拠の信頼性を向上させる．評価実験により正確性と危険予知能力が向上し，実業務への適用可能性を示した．

1 はじめに

NTT 東日本は通信局舎からお客様宅を結ぶ通信ネットワークを構築しており，それらを構成する，光ケーブル，電柱，マンホール，管路など膨大な量の設備を保守している．通信設備の建設，保守の作業は高所作業が多く，安全管理が極めて重要である．安全管理の手法として，作業現場で撮影した画像を基にした危険個所のフィードバックを行っているが，目視確認に依存しており，効率性や品質担保の観点から自動判定する技術が求められていた．

従来の物体検知モデル [1] では特定の物体検出は可能であるが，作業現場全体の状況理解や危険予知といった複雑なタスクには対応できない．近年，Vision-Language Model（VLM）[2, 3, 4] により画像とテキストの統合処理が可能となったが，医療 [5] や建設 [6] の研究が示すように，特定ドメインでは専門知識への適応が不可欠である．

そこで本研究では，構造化知識蒸留と Visual Attention 最適化を統合した，通信設備保守に特化した VLM の構築手法を提案する．データ面では，従来の知識蒸留における汎用 VLM による単純な画像説明に対し，業務マニュアルの組み込み，危険予知の 4 段階構造化，専門用語の正規化により専門知識を効果的に注入する．モデル面では，Soft Registers による Visual Attention Sink 抑制により，背景などの無関係な領域への過剰な注視を軽減し，視覚的根拠の信頼性を向上させる．具体的には，(1) 業務知識を付与した高品質合成データの生成手法，(2) 知識蒸留と LoRA（Low-Rank Adaptation）[7] による効率的ファインチューニング，(3) Soft Registers による Visual Attention 制御，(4) LLM-as-a-Judge[8, 9] によるドメイン特化評価の 4 点を提案する．

2 関連研究

2.1 Vision-Language Models

CLIP[2]，BLIP[10]，LLaVA[3]，Qwen3[4] などの VLM により，画像とテキストの統合処理が可能となった．これらの汎用 VLM は様々なタスクで高性能を示すが，特定ドメインの専門知識には限界がある．

2.2 ドメイン特化型 VLM

医療診断 [5]，建設現場の安全検査 [6]，製造業の外観検査 [11] など，汎用 VLM を特定業界に適用する研究が進められている．通信設備保守では，及川ら [1] がマンホール鉄蓋の種別判定を実現したが，複雑な状況理解には至っていない．本研究は，通信設備保守に特化した VLM の開発手法を提案する．

2.3 知識蒸留

知識蒸留 [12] は，大規模な Teacher モデルの知識を小規模な Student モデルに転移する技術である．Hinton らは Teacher モデルの出力分布（Soft Targets）を活用することで，単なるラベルよりも豊かな情報を Student モデルに伝達できることを示した．近年，Self-Instruct[13] や Orca[14] のように，大規模言語モデルから合成データを生成し，小規模モデルの性能を向上させる手法が提案されている．これらの手法は汎用的なタスクを対象としているが，本研究では業務マニュアルや専門用語を組み込んだ構造化知識蒸留により，ドメイン特化の専門知識を効果的に注入する点が異なる．

2.4 Parameter-Efficient Fine-Tuning

LoRA[7] は，モデルの重み行列に低ランク行列を追加することで，学習パラメータを大幅に削減しながら高性能を維持する手法である．本研究では LoRA を採用し，限られた計算資源で効率的にドメイン適応を実現する．

2.5 LLM-as-a-Judge 評価手法

LLM-as-a-Judge[8, 9] は，人間の評価を模倣した柔軟な評価が可能である．また，llm-jp-eval-mm[15] は日本語視覚言語モデルの自動評価基盤を提供している．従来の ROUGE[16] や BLEU[17] は参照テキストとの表面的な単語一致を評価するため，正解データが存在しないタスクや，専門知識の正確性・危険予知の質といった意味的な評価には適用できない．本研究では，LLM-as-a-Judge 手法を拡張し，通信設備保守の専門知識を考慮した評価基準を組み込む．

3 提案手法

本節では，通信設備保守に特化した VLM の構築手法を述べる．提案手法は，(1) VQA（Visual Question Answering）データセット構築，(2) Visual Attention 制御を統合した LoRA ファインチューニング，(3) LLM-as-a-Judge による評価，の 3 段階から構成される（図 1）．

3.1 VQA データセット構築

約 5 年間の実業務データから 4,567 枚の画像を収集した．これらの画像には質問-回答ペアが付与されていないため，Self-Instruct[13] や Orca[14] のような大規模モデルからの知識蒸留手法を参考に，汎用 VLM（Qwen3-VL-235B-A22B-Instruc）を用いて各画像に対する質問-回答ペアを生成する．本研究の主要な貢献は，業務知識を付与した合成データ生成手法である．従来の知識蒸留では汎用 VLM に単純な画像説明を生成させるが，本手法では以下の 3 つの戦略により専門知識を効果的に注入する（図 2）．

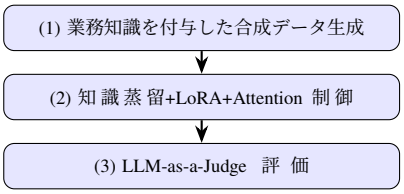

図 1 提案手法の全体フロー．

1. **業務マニュアルの組み込み**: 通信設備保守の安全基準や作業手順をプロンプトに明示的に含める．業務マニュアルの全文は長大でプロンプトに収まらないため，LLM による要約を実施した上で組み込む
2. **危険予知の 4 段階構造化**: (a) 危険認識，(b) 不安全行動特定，(c) 状況詳細化，(d) 危険予知の言語化の段階的な思考プロセスを誘導
3. **専門用語の正規化**: 業務で使用される正確な用語に統一（例：「バケット車」を「高所作業車」に変更）

生成された合成データは専門家レビューを経てデータセットに追加する．

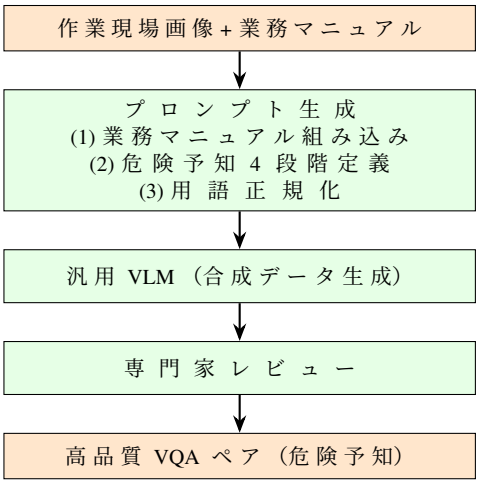

図 2 業務知識を付与した合成データ生成プロセス．

最終的に，各画像に対して 1 件ずつ VQA ペアを生成し，計 4,567 件のデータセットを構築した．

3.2 LoRA によるファインチューニング

ベースモデルとして Qwen3-VL-8B[4]（80 億パラメータ）を使用し，LoRA[7] でファインチューニングする．LoRA はモデルの重み行列 W に低ランク行列 $\Delta W = BA$ を追加することで，元のモデルパラメータを凍結し，追加する低ランク行列のみを学習することで効率的にファインチューニングする．本研究ではランク $r = 16$，$\alpha = 32$ として視覚エンコーダ（Vision Transformer: ViT），aligner，テキストデコーダ（Large Language Model: LLM）の全モジュールに適用した．

3.3 Visual Attention Sink の抑制に向けた検討

VLM のファインチューニングにおいて，画像の背景や無意味な領域を過剰に注視してしまう「Visual Attention Sink（VAS）」現象 [18] が課題となっている．LLM は確率計算の整合性を保つため，文脈的に不要な Attention を吸収する「ゴミ捨て場（Sink）」を必要とするが，適切な Sink が存在しない場合，画像トークンの一部がその役割を担わされ，本来の視覚情報が損なわれる可能性がある [19]．

そこで本研究では，モデルの視線を対象物に正しく誘導するための構造的なアプローチとして，Soft Registers の導入を検討した．具体的には，画像エンコーダ出力と LLM 入力の境界（<|vision_end|> トークン直後）に，学習可能な 8 個のレジスタトークンを挿入し，損失計算を行わない（Masking）ことで，不要な Attention を吸収させる仕組みである（詳細は Appendix B）．

3.4 LLM-as-a-Judge による評価

LLM-as-a-Judge 手法を用いて，通信設備保守の専門知識を考慮した評価基準を設計する．評価基準は，Accuracy（専門用語や安全基準の正確さ），Helpfulness（回答の実用性），Safety（安全基準への適合性），KY（危険箇所の認識と予知の質），Business（業務報告としての適切性）の 5 項目で構成され，各項目を 1–10 点でスコアリングする．評価プロンプトは質問，VLM の回答，評価基準を含み，JSON 形式で評価結果を出力する．

4 実験設定

構築した VQA データセットを訓練データ 2,866 件，検証データ 1,506 件，テストデータ 195 件に分割した．比較モデルとして，(1) ベースラインモデル（Vanilla），(2) Model A（シンプルな合成データ），(3) Model B（提案手法，業務知識を付与した合成データ）の 3 モデルを評価した．専門家が作成した正解データが存在しないため，gpt-5-mini による LLM-as-a-Judge 評価を実施し，Accuracy，Helpfulness，Safety，KY，Business の 5 項目を 1–10 点でスコアリングした．なお，本節で述べる定量評価（Model A，B の比較）には Soft Registers を適用していない通常のモデルを使用した．Soft Registers 導入モデルについては，Visual Attention の改善効果を確認するため，5.3 節および 5.4 節にて別途定性的な評価を行う．

5 実験結果と考察

5.1 評価結果

図 3 に評価結果を示す．各評価基準（Accuracy, Helpfulness, Safety, KY, Business）において，提案手法（Model B）はベースライン（Vanilla）と比較して Accuracy で+5.0%（8.24 点から 8.65 点），KY で+7.6%（8.04 点から 8.65 点），Business で+1.8%（8.00 点から 8.14 点）の改善を達成した．特に KY スコアの大幅な向上は，危険予知の 4 段階構造化により危険認識と不安全行動の特定が促進された結果である．また，Model A と Model B の比較では，業務知識付与により更なる性能向上が確認され（Accuracy は 8.60 点から 8.65 点，Helpfulness は 8.51 点から 8.61 点），単純な画像説明による知識蒸留と比較した提案手法の有効性が示された．全モデルで Safety スコアが 9.2 点以上を維持しており，安全基準への適合性は確保されている．

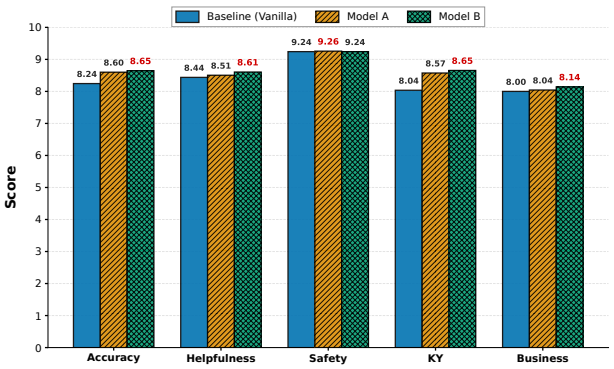

図 3 LLM-as-a-Judge 評価結果の比較．

5.2 考察

Model A（Accuracy が+4.4%，KY が+6.7%向上）の結果は，汎用 VLM による単純な画像説明であっても，合成データによる知識蒸留が一定の効果を持つことを示している．一方，Model B では業務マニュアルの組み込み，危険予知の 4 段階構造化，専門用語の正規化という 3 つの戦略により，専門知識が効果的に注入され，従来の単純な画像説明では得られない業務固有の専門知識と思考プロセスを学習できた．従来の ROUGE[16] や BLEU[17] は表面的な単語一致を評価するため危険予知の質や専門知識の正確性を評価できないが，LLM-as-a-Judge は多面的な品質評価を可能とし，特に KY と Accuracy の 2 軸で安全性と専門性の両面から性能を検証できた．Model B の KY スコア（8.65 点）は実業務における危険予知フィードバックの実現可能性を示しており，今後は実際の作業現場での運用評価が必要である．

5.3 Visual Attention の定性評価

前節までの評価により，提案するデータ構築手法が回答の精度（Accuracy）や危険予知（KY）の質を向上させることが示された．これに加え，実業務における安全管理システムとしての信頼性を担保するためには，モデルが画像の適切な領域を根拠として判断しているか（Grounding）を確認することも重要である．特に，背景などの無関係な領域を過剰に注視してしまう Visual Attention Sink（VAS）現象は，ハルシネーションの一因として知られており [18]，この抑制は正確な危険予知において構造的な課題となる．

そこで本節では，モデルの視覚的注視を物理的に制御するアプローチとして，Soft Registers を導入したモデルにおける Attention Map の定性評価を行った．解析結果を図 4 に示す．

導入前（Baseline）では，質問内容に関わらず画像の左上隅（背景）に強い Attention が発生しており，VAS 現象が顕著に見られた（図 4(a)）．これに対し，Soft Registers を導入したモデルでは，背景への不要な Attention（Sink）が軽減されている様子が確認された（図 4(b)）．さらに，図 4(c) に示すように，Attention が「高所作業中の作業員」等の対象物へ集中的に向く傾向（Object Grounding）が見られた．この結果は，Soft Registers の導入が VLM の視覚的注視を安定化させる効果を持つことを示唆しており，前節で述べたデータ面での改善と合わせることで，より堅牢なシステム構築に寄与する可能性がある．

5.4 本手法の利点と残された課題

本手法（Soft Registers + LoRA）の利点は，計算コストと実用性のバランスにある．Darcet ら [18] が提案した Registers は ViT 全体の再学習を要するが，本手法はごく少数のトークンを追加学習するのみであり，追加コストは軽微である．これにより，計算資源が限られる現場のオンプレミス環境等でも，視覚的注視の制御によるハルシネーション抑制効果を享受できる．

一方で，課題も確認された．図 4(c) を詳細に観察すると，左上の Sink は解消されたものの，右上領域に新たな微弱な Attention の集中（Secondary Sink）が発生している．これは，少数のレジスタトークンですべての不要な Attention を吸収しきれなかった場合，モデルが再び画像の隅を Sink として利用しようとする現象であると考えられる．実運用に向けては，レジスタトークン数の最適化や，Attention の分散をさらに制御する学習手法の検討が必要である．

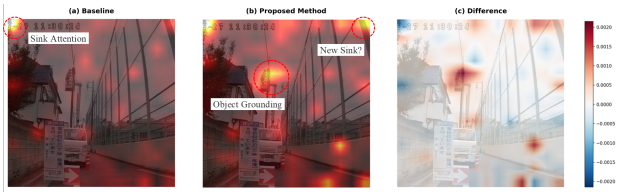

図 4 Visual Attention 比較．(a)Baseline，(b) 提案手法（Sink 抑制・Grounding），(c) 差分．

6 まとめ

本研究では，業務知識を付与した合成データ生成による知識蒸留に基づく，通信設備保守に特化した VLM の開発手法を提案した．業務マニュアルの組み込み，危険予知の 4 段階構造化，専門用語の正規化という 3 つの戦略により，汎用 VLM から専門知識を効果的に抽出し，ドメイン特化型 VLM に注入する．評価実験により，Accuracy で+4.9%，KY で+7.7%の改善を達成し，実業務での危険予知フィードバックシステムの実現可能性を示した．

今後の課題として，実際の作業現場での運用評価，他のドメインへの適用可能性の検証，および Visual Attention Sink（VAS）抑制効果と精度向上の因果関係の定量的検証が挙げられる．

参考文献

[1] 及川大輝, 勝村玲音, 和田雅樹, 島原広季, 相原貴明. 深層学習を用いたマンホール鉄蓋種別判定および不良箇所検出にむけた検討. 人工知能学会第 35 回全国大会論文集, 2021.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In **ICML**, 2021.

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **NeurIPS**, 2023.

[4] Qwen Team. Qwen3 technical report. **arXiv preprint arXiv:2505.09388**, 2025.

[5] Chunyuan Li, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In **NeurIPS**, 2023.

[6] Xuezheng Chen and Zhengbo Zou. Are large pre-trained vision language models effective construction safety inspectors? **arXiv preprint arXiv:2508.11011**, 2025.

[7] Edward J Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In **ICLR**, 2022.

[8] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In **NeurIPS**, 2023.

[9] 中山功太, 児玉貴志, 鈴木久美, 宮尾祐介, 関根聡. llm-jp-judge: 日本語 LLM-as-a-Judge 評価ツール. 言語処理学会第 31 回年次大会, 2025.

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In **ICML**, 2022.

[11] 岡部健太, 遠藤隆夫, 石上将太郎, 中村光貴, 仁平雅也, 乙村浩太郎, 羽藤淳平. VLM を用いたドメイン特化生成画像の定量評価. 言語処理学会第 31 回年次大会, 2025.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. **arXiv preprint arXiv:1503.02531**, 2015.

[13] Yizhong Wang, et al. SELF-INSTRUCT: Aligning language models with self-generated instructions. In **ACL**, 2023.

[14] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. **arXiv preprint arXiv:2306.02707**, 2023.

[15] 前田航希, 杉浦一瑳, 小田悠介, 栗田修平, 岡崎直観. llm-jp-eval-mm: 日本語視覚言語モデルの自動評価基盤. 言語処理学会第 31 回年次大会, 2025.

[16] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, 2004.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In **ACL**, pp. 311–318, 2002.

[18] TimothÃ©e Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In **ICLR**, 2024.

[19] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In **ICLR**, 2025.

A　LoRA ファインチューニングパラメータ

本研究で使用した LoRA ファインチューニングの主要なパラメータを表 1 に示す．

表 1　LoRA ファインチューニングの主要パラメータ

パラメータ	値
Model	Qwen3-VL-8B
Training type	LoRA
LoRA rank (r)	16
LoRA alpha (α)	32
LoRA dropout	0.05
Target modules	all-linear
Freeze (LLM/ViT/align.)	false
Data type	bfloat16
Epochs	10
Batch size	4
Grad. accumulation	4
Effective batch size	16
LR (LLM)	1×10^{-4}
LR (ViT)	1×10^{-5}
LR (aligner)	1×10^{-5}
LR scheduler	cosine
Warmup ratio	0.1
Weight decay	0.01
Max length	8192

B　Visual Attention Sink の解析詳細

Soft Registers の物理的な配置構造を図 5 に示す．画像エンコーダ（ViT）の出力直後かつテキストトークンの直前に，学習可能なレジスタトークンを挿入することで，画像特徴量を破壊することなく不要な Attention を吸収させる設計となっている．

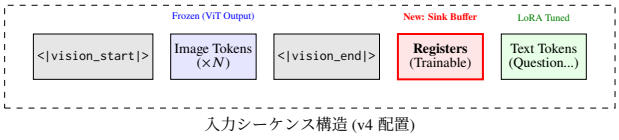

図 5　Soft Registers の導入位置．画像処理終了トークンの直後にレジスタトークンを挿入する．

B.1　実装詳細

追加した 8 個のレジスタトークンは，Qwen3-VL-8B の既存の埋め込み層の統計量（平均・分散）を用いて初期化し，LayerNorm を適用せずに学習した．学習率は LLM 本体（LoRA，1×10^{-4}）よりも高い 1×10^{-3} に設定し，早期に Sink としての機能を獲得させる工夫を行った．