日本語特許を利用した画像キャプションデータセットの構築と段階的情報抽出によるキャプション生成手法の提案

喜多俊介[1]　森辰則[1]　小野寺理恵[2]　伊藤拓海[2]
[1] 横浜国立大学大学院　[2] 株式会社 IHI
kita-shunsuke-yz@ynu.jp tmori@ynu.ac.jp onodera3892@ihi-g.com ito9762@ihi-g.com

概要

近年、社内文書の電子化と利活用が進む一方、解析基盤はテキスト中心であり、技術文書で重要となる図表の情報を十分に活用できていない。そこで本研究では、図面単位で画像と画像外の本文テキストを扱える日本語公開特許データセットを構築するとともに、短文と長文の 2 種のキャプション生成タスクを設計した。さらに、テキスト選択、情報抽出、キャプション生成を単一プロンプト内で段階的に実行させる Chain-of-Selection-Extraction-Captioning（CoSEC）を提案する。評価の結果、短文では入力するテキスト情報を絞る戦略が、長文では十分なテキスト情報を与えた上で CoSEC により要点へ注意を誘導する方針が有効であることが示唆された。

1　はじめに

近年、企業の DX により社内文書の電子化と利活用が進む一方、解析対象は依然としてテキスト中心である。例えば我々の事故事例文書絞り込み検索システム [1] もテキストのみを扱うが、技術文書では図表が情報の核心を担う場合が多く、テキストのみの解析では十分な知見を得にくい。画像情報を既存のテキスト解析基盤へ統合する手段として画像からのキャプション生成が有効だが、汎用キャプション生成技術は日常画像を主対象として発展してきたため、線画・構造図が多い技術文書では専門用語や図中のラベル（参照用文字列）との整合が難しく、重要情報の欠落やハルシネーション [2] が生じやすい。

そこで本研究では、技術文書の例として日本語公開特許公報を対象に、図面単位で画像と画像外の本文テキスト（以下、画像外テキスト）を扱える日本語公開特許データセットを構築し、これに基づく画像キャプション生成タスクを設計する。さらに、図面に対応する本文テキストを併用した生成手法を検討し、技術文書における画像由来情報を既存のテキスト解析基盤へ統合可能とする基盤整備を目指す。

2　関連研究

2.1　特許文書データセット

特許文書は書式が規定され大量に無償公開されており、さらに本文中で図面内容が明示的に記述されることが多いため、図面画像と説明文のペアを比較的容易に構成でき、画像内容理解の評価用データソースとして広く用いられる [3, 4]。代表例として PatFig[3] と PatentDesc-355K[4] がある。

PatFig は欧州特許出願の約 3 万枚の図面に短文・長文キャプションといくつかのメタデータを付与したデータセットであり、PatentDesc-355K は米国特許文書の約 35.5 万枚の図面へ短文・長文キャプションを付与したデータセットである。いずれも短文は特許文書中の「Brief Description of Drawings」セクションから抽出し、長文については、PatFig は本文中で同一図番号が連続参照される箇所を抽出して 50–500 トークンに収まるものを採用し、PatentDesc-355K は当該図を参照する箇所を連結した上で複数図同時参照を除外し、最大 500 トークンに切り抜いている。

2.2　画像キャプション生成

画像キャプション生成（Image Captioning）は、入力画像に基づいて自然言語キャプションを生成するタスクである。一方、主要ベンチマークは日常画像が中心であり [5]、構造図等が多い技術文書では専門用語や参照符号との整合が難しく、重要要素の欠落やハルシネーションが課題とされている [3, 4]。

この課題に関連して、News Image Captioning ではニュース画像と記事テキストを入力とし、固有名詞や状況説明など画像単独では得にくい情報を記事から補完する設定が検討されている。記事本文は長文

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

であるため、画像との関連度が高い文 [6] やタイトル・概要 [7] などを選択して用いることで品質向上が示され、テキスト選択から生成までを一体として学習する手法も提案されている [8]。

2.3 本研究の貢献

本研究の貢献は以下のとおりである。

- PatFig[3]、PatentDesc-355K[4] を踏まえ、日本語公開特許公報を対象に、図面画像と画像外テキストを図面単位で対応付ける日本語公開特許データセット構築プログラムを作成し、それを基にキャプション生成タスクを設計する。
- News Image Captioning におけるテキスト選択から生成までを一体として扱う枠組み [8] を踏まえ、選択 → 抽出 → 生成を単一プロンプトで段階的に実行する CoSEC を提案する。
- 実験を通じて短文では入力テキストの絞り込み、長文では十分な文章量＋焦点化（CoSEC）が有効になり得ることを示す。

3 日本語公開特許データセット

本研究では、特許情報のダウンロードサービス [9] で取得した日本語公開特許公報 PDF を基に、日本語公開特許データセットを構築するプログラムを作成した。表 1 に各フィールドの説明を示す。

表 1 レコードの各フィールドの説明

フィールド名	型	説明
id	str	レコードの一意な識別子
image_path	str	図面画像ファイルの相対パス
pub_number	str	公開特許公報の公開番号
caption	dict	図面の参照キャプション
→short	str	短文参照キャプション
→long	str	長文参照キャプション
text_context	dict	公開特許公報のテキスト情報
→title	str	「発明の名称」セクション
→abstract	str	「要約」セクション
→claims	str	「特許請求の範囲」セクション
→terms	dict	参照符号と用語の対応辞書（「符号の説明」セクション）
→description	str	「発明の詳細な説明」セクション
fig	dict	図面のメタデータ
→index	int	文書中の図番号
→total	int	文書内の総図数
→label	str	図ラベル（例：「図 1」）

3.1 構築方法

本データセットでは、1 レコードを 1 枚の図面画像に対応付ける。1 公報（pub_number）から複数レコードを生成し、公報内では図番号（fig.index）で図面を区別する。id は pub_number を西暦 4 桁+6 桁の 10 桁に正規化した番号と fig.index から一意に定まり、id = {10 桁番号}_fig{fig.index}とする。

図面画像は公報 PDF に埋め込まれた画像を抽出して PNG で保存する。保存時にアスペクト比を維持して長辺を 1024px に縮小し、不足分を白色パディングして画像サイズを 1024 × 1024 に統一する。

短文・長文の参照キャプション（データセットにおける「正解」に対応）は先行研究 [3, 4] を参考に作成する。短文は公報内の「図面の簡単な説明」セクションから該当図の説明を抽出し、文末表現やメタ表現（例：「である。」「本発明」）をルールベースで削除する。長文は公報内の「発明の詳細な説明」セクションにおける当該図の連続参照範囲を抽出し、そのテキストと短文参照キャプション、画像を入力として MLLM（GPT-5[10]）で生成する。先行研究ではトークン数に基づいて文量を調整しているが、本研究では MLLM により文量を制御しつつ、画像とテキストを同時参照して無関係記述の混入抑制を図る。ハルシネーションが混入する可能性は残るが、長文キャプション生成において近年の MLLM が人手と同等以上の性能を示す報告もある [11] ため、本研究ではこのアプローチを採用する。

画像外テキスト（title/abstract/claims/terms/description）は、公報 PDF からセクション見出しに基づいてルールベースで抽出して作成する。

3.2 タスク設定

本研究では日本語公開特許データセットを利用し、日本語特許文書を対象としたキャプション生成タスクを設計した。図 1 にその概略図を示す。

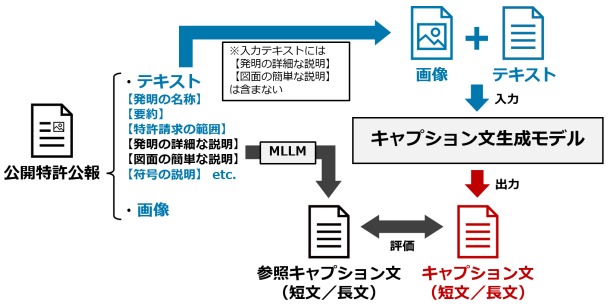

図 1 タスクの概略図

特許文書では、画像内容の説明に適切な画像外テキストを活用することで、生成されるキャプションの品質が向上することが示されている [3]。よって本タスクでは、入力として図面画像に加えて title、abstract、terms、claims の画像外テキストを用いる。なおこの際、入力とする画像外テ

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

キストが適切でない場合はモデルに余計な情報を与えてしまうこととなり、生成されるキャプションの品質が低下してしまうことも報告されている [3]。したがって本タスクでは、入力とする画像外テキストの種類やその活用方法も重要となる。ただし、description は図面説明として最も直接的であり、入力に含めると参照文の再生成に近い設定となるため、本タスクでは汎用的な付随情報（title/abstract/claims/terms）のみでどこまで説明可能かを評価することとする。

また本タスクは、画像内容を簡潔に説明した短文キャプションまたは詳細に説明した長文キャプションのいずれかを出力とし、生成キャプションが参照キャプションと同等の内容を記述できているかを性能目標とする。なお、参照キャプションは「発明の詳細な説明」セクションにおける当該図の説明箇所に基づいて作成されるため、画像と本文の対応が明確な基準文として用いることができる（付録参照）。

4 Chain-of-Selection-Extraction-Captioning（CoSEC）

技術文書では画像に対応する説明が本文中に散在し、無関係情報も多い。そこで本研究では、テキスト選択、情報抽出、キャプション生成を単一プロンプト内で段階的に実行させた上で、最終キャプションを生成する CoSEC を提案する（図 2）。本手法は、News Image Captioning においてテキスト選択から生成までを一体として扱う枠組み [8] と、中間的な推論ステップを示した入出力例をプロンプトとして与える Chain-of-Thought Prompting（CoT）[12] に着想を得ている。CoSEC は以下の 3 段階からなる。

1. **Selection**: 入力テキストから画像関連の文・節のみを選択する。
2. **Extraction**: 固有表現や関係を抽出し、後段で利用しやすい形に整理する。
3. **Captioning**: 画像と抽出要素を根拠として整合的なキャプションを生成する。

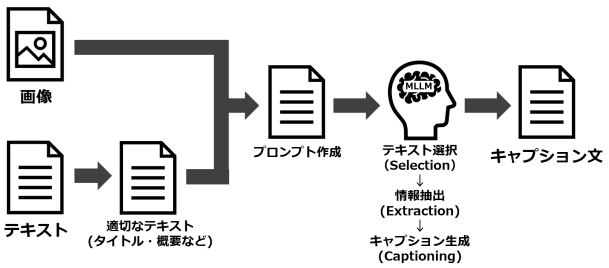

図 2 Chain-of-Selection-Extraction-Captioning の概略図

先行研究である推論過程の明示を主目的とする CoT 型キャプション生成 [13] とは異なり、画像外テキスト活用を選択 → 抽出 → 生成の段階処理として単一のプロンプトで実行させる点に特徴がある。

5 評価実験

日本語特許文書を対象としたキャプション生成タスクにおける提案手法（CoSEC）の有効性と、入力とする画像外テキスト（title/abstract/claims/terms）の設計が生成に与える影響を検証する。

5.1 実験設定

表 2 に示す補助要素の有無が生成に与える影響を検証する。ベースラインは画像のみ入力（Base）とし、Base に対して各補助要素を 1 つずつ追加した条件を比較する。加えて、文章量が大きい claims については表 3 の処理を適用し、処理の違いによる効果も評価する。さらに、これらの処理を単一プロンプトで段階的に実行させる CoSEC（NER/NER+RE）が生成に与える影響を評価する（NER と NER+RE の処理内容は表 3 参照）。最後に、良好だった補助要素・処理を組み合わせた条件（Multiple）を追加し、複数テキスト併用の効果を分析する。

表 2 評価実験で利用する補助要素の説明

補助要素	説明
title	title を画像外テキストとして利用する
abst	abstract を画像外テキストとして利用する
term	terms を画像外テキストとして利用する
claim	claims を画像外テキストとして利用する

表 3 評価実験で claims に対して行う処理の説明

処理	説明
Select	claims から関連箇所を選別して利用する
NER	claims から関連箇所を選別し、そこから固有表現を抽出して利用する
NER+RE	claims から関連箇所を選別し、そこから固有表現を抽出する。その後、抽出した固有表現間の関係性を示す短文を抽出して利用する

評価用データセットは日本語公開特許データセット構築プログラムにより構築した。特許情報の一括ダウンロードサービスで取得可能であった 2025 年 10 月 29 日更新分バルクデータから、選択基準に基づいて人手で選択した公開特許公報 30 件に含まれる画像 238 枚を収集し、重複のない無作為抽出で 100 枚を選定した。選択基準は、(i) 製造・装置・プログラムに関する発明であること、(ii) 文書構造が本タスク設定に適合すること、の 2 点である。

MLLM は API 経由で GPT-5 を用い、API リクエスト設定は reasoning_effort を"high"とした。

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 4　評価実験結果の抜粋

	短文					長文				
条件	METEOR	CIDEr-D	JaSPICE	RefPAC-S++	GPT-4.1	METEOR	CIDEr-D	JaSPICE	RefPAC-S++	GPT-4.1
Base	0.176	0.717	0.323	0.587	0.535	0.269	0.433	0.483	0.586	0.600
+term	**0.258**	**1.693**	**0.462**	**0.600**	0.565	0.286	0.480	0.533	**0.601**	**0.635**
+claim	0.222	1.113	0.393	0.591	0.555	0.286	0.532	0.533	0.595	0.621
NER	0.238	1.423	0.425	0.595	**0.571**	0.277	0.451	0.520	0.590	0.604
NER+RE	0.230	1.224	0.411	0.592	0.566	0.282	0.542	0.521	0.597	0.623
CoSEC(NER)	0.222	1.110	0.381	0.590	0.552	0.290	0.563	0.533	0.593	0.623
CoSEC(NER+RE)	0.217	1.165	0.374	0.590	0.545	0.291	0.578	0.536	0.597	**0.635**
Multiple	0.252	1.447	0.432	0.590	0.548	**0.297**	**0.600**	**0.549**	0.597	0.631

5.2　評価手法

評価には参照との類似性に基づく指標である METEOR[14]、CIDEr-D[15]、JaSPICE[16]、RefPAC-S++[17] と、VLM-as-a-Judge として GPT-4.1-as-a-Judge を用いる [11]。METEOR と CIDEr-D は MeCab[18] と NEologd[19] で分かち書きした上で、Microsoft COCO Captions[5] の評価ツール（pycocoevalcap）を用いて算出した。RefPAC-S++は英語を主対象としているため、評価対象文・参照文を GPT-5 Nano[10] で英訳した。また、RefPAC-S++の算出時にはデータセットの構築時に行う画像の白色パディングがスコアに悪影響を与える可能性があるため、パディング無し画像（長辺のみ 1024px に縮小）を別途用いた。GPT-4.1-as-a-Judge は、GPT-4.1[20] に図面画像、参照キャプション、生成キャプションを入力し、内容整合（Consistency）と文章品質（Fluency）の 2 観点を A/B/C/D で判定させた。その後、各ラベルの出力候補確率に対し、$A/B/C/D \mapsto 1.0/0.7/0.3/0.0$ の写像による期待値を取り、$[0, 1]$ の連続値スコアへ変換した上で、2 観点の平均を最終スコアとした。なお、各観点とラベル定義の詳細は付録を参照されたい。

5.3　結果と考察

評価実験結果の抜粋を表 4 に示す。なお、評価実験における全条件の結果は付録を参照されたい。

画像外テキストの付与は短文・長文の両設定で Base を上回り、生成品質の改善に寄与した。短文では `terms` が一貫して良好であり、先行研究 [3] で指摘された悪影響は、本実験では MLLM が必要語を選別できたため生じにくかった可能性がある。長文でも補助要素の追加は有効で、`terms` と `claims` の改善が大きく、画像外テキストの文章量の増加により具体的な記述が得られやすいことが示唆される。

`claims` の処理は短文では入力圧縮が有効で、とくに NER が良好であった一方、NER+RE の利得は限定的で、抽出結果がテンプレート的になり生成の自由度を制約した可能性がある。長文では情報削減が不利になり得るが、NER+RE が一部指標で改善し、不要情報の除去と必要要素の保持のバランスが取れた場合に有効となり得ることが示唆される。

CoSEC は短文では不利で、元テキストを保持した段階処理により入力が増え不要情報が混入しやすいと考えられる。一方、長文では CoSEC が有効で、要点（固有表現・関係）への注意誘導により脱線や冗長化を抑えつつ情報を保持できた可能性がある。

複数テキストの併用は短文では入力増による不要情報混入を招きやすく、`terms` のみを与える条件の方が良好な結果となった。対照的に長文では複数要素の併用がいくつかの指標で最良となった。

総じて、短文では少量手掛かりを優先して入力を絞り、長文では十分な文章量を与えた上で焦点化（例：CoSEC）を行う設計が有効と考えられる。

6　おわりに

本研究では、日本語公開特許公報 PDF の図面画像と画像外テキストを扱える日本語公開特許データセットを構築し、短文・長文の画像キャプション生成タスクを設計した。さらに、選択 → 抽出 → 生成を単一プロンプトに組み込む CoSEC を提案した。

評価の結果、画像外テキストの付与は短文・長文とも品質を改善し、短文では入力する画像外テキストを絞る戦略が有効であり、長文では十分なテキスト情報を与えた上で CoSEC により要点へ注意を誘導する方針が有効であることが示唆された。

今後は、データセット構築における MLLM 依存工程の効率化、短文・長文それぞれに適した画像外テキストの選択戦略の高度化、およびデータ規模・分野の拡大や人手評価による多面的評価を通じて、一般化可能性と実運用性を検証する必要がある。

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] 福岡康大, 大八木悠聖, 喜多俊介, 深草理貴, 森辰則, 伊藤拓海, 小野寺理恵. 事故事例文書絞り込み検索システムの構築. 言語処理学会第 31 回年次大会発表論文集, pp. 3217-3221, 2025.

[2] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, vol. 43, no. 2, article 42, 2025.

[3] Dana Aubakirova, Kim Gerdes, Lufei Liu. PatFig: Generating Short and Long Captions for Patent Figures. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 2843-2849, 2023.

[4] Shreya Shukla, Nakul Sharma, Manish Gupta, Anand Mishra. PatentLMM: Large Multimodal Model for Generating Descriptions for Patent Figures. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 19, pp. 20488-20496, 2025.

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv preprint arXiv:1504.00325 [cs.CV], 2015.

[6] Anwen Hu, Shizhe Chen, Qin Jin. ICECAP: Information Concentrated Entity-aware Image Captioning. In Proceedings of the 28th ACM International Conference on Multimedia (MM ’ 20), pp. 4217–4225, 2020.

[7] Mingyang Zhou, Grace Luo, Anna Rohrbach, Zhou Yu. Focus! Relevant and Sufficient Context Selection for News Image Captioning. In Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 6078–6088, 2022.

[8] Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Xiaojun Wan. EAMA: Entity-Aware Multimodal Alignment Based Approach for News Image Captioning. arXiv preprint arXiv:2402.19404 [cs.CV], 2024.

[9] 特許庁. 特許情報の一括ダウンロードサービスについて. https://www.jpo.go.jp/system/laws/sesaku/data/download.html, 2025. [accessed: 2025-11-05]

[10] OpenAI. GPT-5 System Card. https://openai.com/ja-JP/index/gpt-5-system-card/, 2025. [accessed: 2025-12-12]

[11] Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, Jiajun Chen. CapArena: Benchmarking and Analyzing Detailed Image Captioning in the LLM Era. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 14077-14094, 2025.

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022), 2022.

[13] Kohtaro Tanaka, Kohei Uehara, Lin Gu, Yusuke Mukuta, Tatsuya Harada. Content-Specific Humorous Image Captioning Using Incongruity Resolution Chain-of-Thought. In Findings of the Association for Computational Linguistics: NAACL 2024, pp. 2348-2367, 2024.

[14] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72, 2005.

[15] Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566-4575, 2015.

[16] Yuiga Wada, Kanta Kaneda, Komei Sugiura. JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models. In Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pp. 424-435, 2023.

[17] Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara. Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training. International Journal of Computer Vision, vol. 133, pp. 7647–7671, 2025.

[18] Taku Kudo. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. https://taku910.github.io/mecab/, 2005. [accessed: 2025-10-06]

[19] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会発表論文集, pp. 875-878, 2017.

[20] OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, 2025. [accessed: 2025-12-17]

[21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7514-7528, 2021.

[22] Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR 162:12888-12900, 2022.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A　補足情報

表 5　GPT-4.1-as-a-Judge におけるキャプションの評価観点

観点	説明（A/B/C/D）
Consistency	**A:** 参照キャプションの主要内容（対象・主要構成要素・相互関係・必要に応じて動作/流れ）を過不足なく表現し、かつ画像と矛盾しない。画像から裏付け不能な断定がなく、参照キャプションとの食い違いもない（言い換えは可）
	B: 全体として画像・参照キャプションの両方と概ね整合するが、軽微な欠落/過剰詳細/曖昧さ/軽い誤解がある。ただし主要内容は維持され、画像説明として致命的な誤解は生まない
	C: 参照キャプションの重要要素・関係の欠落が大きい、または参照キャプションにない内容の混入が目立つ、あるいは画像と整合しない記述/裏付け不能な断定が含まれる（主要内容の一部が崩れ、信頼性が低い）
	D: 参照キャプションとほぼ一致せず別内容になっている、または主要内容が画像と矛盾する/ハルシネーションが支配的で、画像・参照キャプションに対する整合性評価として成立しない
Fluency	**A:** 日本語として自然で文法誤りがなく読みやすい。指示語・省略・接続が適切で、文内の対象や関係が一貫し、矛盾や飛躍がない（冗長な定型句や同義反復が少ない）
	B: 多少不自然・冗長な箇所はあるが意味は明確に通る。概ね一貫するが、指示の曖昧さや説明順の小さな飛びが一部ある
	C: 文法・語彙・係り受けの問題や冗長さが複数あり読みづらい。対象の揺れや参照関係の不明瞭さ、論理の飛躍が目立ち理解に負荷がかかる
	D: 不自然さが顕著で意味が取りにくい/文として崩れている。矛盾や一貫性の破綻により文章として成立しない

表 6　評価実験結果

	短文					長文				
条件	METEOR	CIDEr-D	JaSPICE	RefPAC-S++	GPT-4.1	METEOR	CIDEr-D	JaSPICE	RefPAC-S++	GPT-4.1
Base	0.176	0.717	0.323	0.587	0.535	0.269	0.433	0.483	0.586	0.600
+title	0.241	1.382	0.428	0.590	0.562	0.272	0.481	0.507	0.593	0.626
+abst	0.236	1.251	0.420	0.591	0.561	0.281	0.473	0.519	0.588	0.625
+term	**0.258**	**1.693**	**0.462**	**0.600**	0.565	0.286	0.480	0.533	**0.601**	**0.635**
+claim	0.222	1.113	0.393	0.591	0.555	0.286	0.532	0.533	0.595	0.621
Select	0.232	1.335	0.416	0.593	0.566	0.282	0.494	0.530	0.595	0.610
NER	0.238	1.423	0.425	0.595	**0.571**	0.277	0.451	0.520	0.590	0.604
NER+RE	0.230	1.224	0.411	0.592	0.566	0.282	0.542	0.521	0.597	0.623
CoSEC(NER)	0.222	1.110	0.381	0.590	0.552	0.290	0.563	0.533	0.593	0.623
CoSEC(NER+RE)	0.217	1.165	0.374	0.590	0.545	0.291	0.578	0.536	0.597	**0.635**
Multiple	0.252	1.447	0.432	0.590	0.548	**0.297**	**0.600**	**0.549**	0.597	0.631

B　予備実験

本データセットでは長文参照キャプションを MLLM で自動作成するため、ハルシネーション等による画像との不整合が品質へ与える影響を予備的に確認した。公開特許公報 16 件由来の図面 87 枚を対象に、GPT-5（`reasoning_effort="high"`）で長文を生成し、参照なし指標 CLIPScore[21] と BLIP[22] の Image-Text Matching(ITM) スコアを用いた指標（以下 BLIPScore と呼ぶ）で画像整合性を評価した。長文の入力長制約に対応するためキャプションをスライディングウィンドウ（隣接窓の重なりは 20 トークン）で分割し、チャンクスコアを max/mean で集約した。両指標が英語を主対象とするため、キャプションは GPT-5 Nano で英訳して算出した。結果（表 7）より、抽出テキストは CLIP-max が高い一方で BLIPScore が低く、画像関連語を含むが冗長・不整合な記述も混在する可能性が示唆された。これに対し生成長文は BLIPScore が最大であり、関連情報の取捨選択と整形により画像整合性が高い記述になっていると考えられる。また、短文参照キャプションと生成長文キャプションを比較すると、CLIP-max および BLIPScore（max/mean）は生成長文キャプションが上回った。一方で CLIP-mean についても両者は概ね同等であった。以上より、生成長文キャプションは、短文参照キャプションと同程度以上の画像整合性を有することが示唆される。以上の結果より、長文参照キャプションの自動作成がデータセット品質へ与える悪影響は限定的である可能性がある。

表 7　長文参照キャプションの評価結果

	CLIP-max	CLIP-mean	BLIP-max	BLIP-mean
短文参照キャプション	0.262	**0.262**	0.659	0.659
抽出テキスト	**0.286**	0.258	0.449	0.353
生成長文キャプション	0.278	0.261	**0.690**	**0.690**

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).