CMDR: 文脈を考慮したマルチモーダル文書検索

田中涼太　長谷川拓　西田京介
NTT 株式会社　人間情報研究所
{ryota.tanaka, taku.hasegawa, kyosuke.nishida}@ntt.com

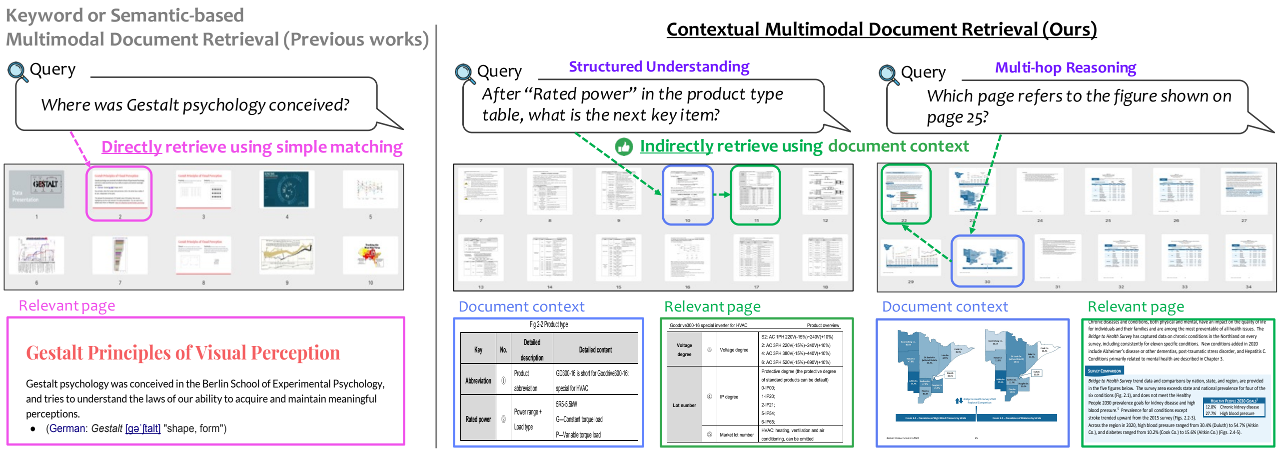


図 1　従来研究と CMDR-Bench の比較．CMDR-Bench は，文脈情報を基にクエリに関連するページを検索する．

概要

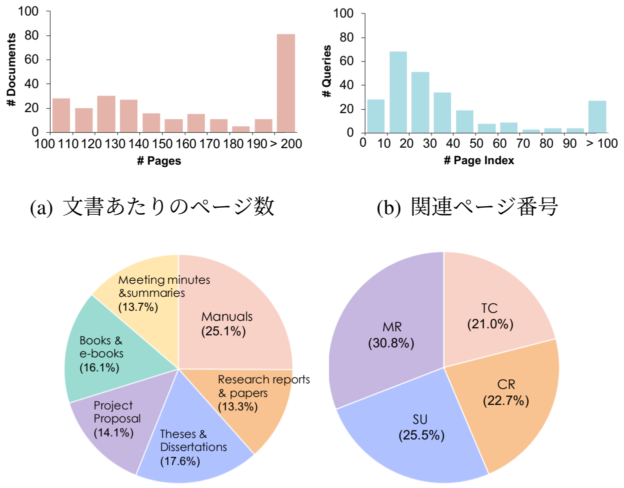
複数ページに跨る文脈を考慮したマルチモーダル文書検索タスク CMDR およびベンチマーク CMDR-Bench を提案する．CMDR-Bench は，文書全体の構成や依存関係といった文脈情報を手掛かりとして，クエリに関連するページを特定する高度な検索を要求する．また，複数ページを同時に埋め込み，文脈情報を統合的に表現する新たな検索モデル CMDR-Embed を提案する．さらに，同一文書内での識別性を保ちながら文脈を活用する対照学習 CMCL を導入する．実験により，文脈情報を活用しない従来手法に対して，提案手法はインデックス化を高速にしつつ，最も高い検索性能を達成した．

1　はじめに

我々が扱う文書はテキストや視覚情報（図表など）を含み，多様な種類・形式が存在する．こうした多様な文書に対して，ユーザの要望を満たす情報を正確に検索する技術の実現は，AI 分野及び産業分野における重要課題の一つである．この課題に対し，テキストのみを扱う従来の検索技術 [1, 2] とは異なり，文書を画像として扱い検索を行う**マルチモーダル文書検索** [3, 4, 5, 6, 7] が有望視されている．

マルチモーダル文書検索における従来データセット [3, 4] では，図 1 に示すように，クエリとページの内容に基づく単純なテキスト・意味一致を問うタスクに留まっており，ページに跨る内容や文書全体の構成・章立てなどの**文脈**（Context）を手掛かりとした検索能力を評価できていない．さらに，従来手法 [5, 6, 7] では，複数ページからなる文書をページ単位で独立にエンコードするため，ページ間に跨る情報依存が存在しないことを暗黙に仮定しており，文脈を考慮した検索手法は未だ実現されていない．

本研究では，複数ページに跨る文脈を考慮した新たなマルチモーダル文書検索タスク **CMDR**（**C**ontextual **M**ultimodal **D**ocument **R**etrieval），およびベンチマーク **CMDR-Bench** を提案する．図 1 に示すように，CMDR-Bench はページを跨いだ複雑な推論を必要とする検索ベンチマークである．さらに，複数ページを同時に埋め込み化することで，文脈を考慮する新たなモデル **CMDR-Embed** を提案する．CMDR-Embed は，文脈情報の活用と，同一文書内における識別性を両立させる新たな対照学習 **CMCL**（**C**ontextual **M**ultimodal **C**ontrastive **L**earning）により学習される．実験の結果，従来の文脈情報を活用しない検索モデルに対して，インデックス化にかかる時間を削減しつつ，大幅な性能向上を達成した．



(a) 文書あたりのページ数　(b) 関連ページ番号

(c) 文書カテゴリ　(d) クエリカテゴリ

図 2　CMDR-Bench における文書とクエリの分布.

2　CMDR-Bench

2.1　CMDR タスク

クエリ q と N 枚のページからなる文書 (画像系列) $\mathbf{I} = \{I_1, \ldots, I_N\}$ が与えられたとき，文書の文脈情報を活用しながら，q に関する回答を含む関連ページを検索する．CMDR における関連ページは，文脈を介してクエリと結び付けられている (図 1 参照).

2.2　データ収集

文書収集　Common Crawl と ManualsLib から多様な PDF 文書を収集し，ページ数（100 ページ以上）および言語（英語）によってフィルタリングを行った．その後，得られた文書を 6 種類に手動で分類し，各ページを画像としてレンダリングした．

クエリ作成　以下の 4 種類のクエリタイプを設計し，各文書につき 3 件以上人手でクエリを作成した．(1) **Text Completion (TC)**：テキスト内容が複数ページに分割されており，ページに跨った意味理解の必要がある．(2) **Coreference Resolution (CR)**：参照表現を他ページに存在する文脈情報を基に理解する必要がある．(3) **Structured Understanding (SU)**：複数ページに分散する構造情報（例：表）を正しく理解する必要がある．(4) **Multi-hop Reasoning (MR)**：離れたページに存在する複数の情報を連鎖的に結び付ける必要がある．

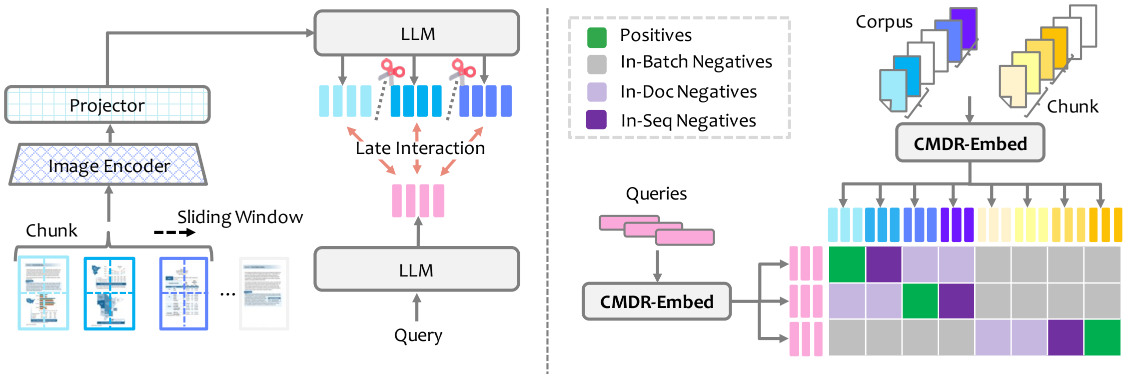
品質管理　最先端の検索モデルを用いて，人手による品質管理を支援した．具体的には，文脈情報を活用しない 3 件の検索モデル [8, 9, 5] を用いて検索し，各クエリに対する Recall@1 の平均値を算出した．スコアが 1.0 のクエリは，文脈情報を利用せずとも解決可能なケースである可能性が高い．そのため，ページ間の依存関係をより強く必要とするように，これらのクエリを修正した．一方で，スコアが低いクエリについては，アノテーション誤りに起因するものかを精査し，誤りが確認されたケースは修正または除外した．最終的に，全クエリのうち 16% が除外され，4% が修正された．

表 1　関連データセットとの比較．Query annotation: M (Manual), S (Synthesized), R (Repurposed).

Benchmarks	Require Context	Multi Modal	Multi-hop Type	Query Annot.	#Avg. Pages
ConTEB [10]	✓	✗	–	R	1.0
MP-DocVQA [11]	✗	✓	–	R	8.1
SlideVQA [3]	✗	✓	Direct	M+R	20.0
MMLB-Doc [4]	✗	✓	Direct	M+R	47.5
LongDocURL [12]	✗	✓	Direct	S	85.6
MMDocIR [13]	✗	✓	Direct	R	65.1
ViDoRe [5]	✗	✓	–	S+R	1.0
OpenDocVQA [7]	✗	✓	Direct	S+R	3.1
VisR-Bench [14]	✗	✓	–	S	4.5
CMDR-Bench (Ours)	✓	✓	Indirect	M	183.5

2.3　統計情報と関連研究との比較

図 2 に示すように，CMDR-Bench は，6 種類の文書タイプに渡る 255 件の長文書に対して，計 800 件のクエリを含む．表 1 に，CMDR-Bench と関連データセットとの比較を示す．**CMDR-Bench は，文脈理解とマルチモーダル文書検索を同時に要求する初のベンチマークである**．既存のマルチモーダル文書検索 [11, 3, 4, 13, 7, 15, 14] は複数ページ文書を扱うものの，主にテキスト一致に基づく *Direct* な検索に焦点を当てている．これに対し，CMDR-Bench は，文脈情報を活用して，クエリに明示的に現れない関連ページを特定する *Indirect* なマルチモーダル検索を対象とする点で独自である．**CMDR-Bench の全クエリは人手で付与されており，高品質かつ実世界の利用状況をより忠実に反映する**．一方，従来研究は，合成（例：VisR-Bench [14]）や再利用（例：ConTEB [10]）に依存しており，検索に適さないクエリが含まれる．さらに，**CMDR-Bench は極めて長い文書を対象とし，困難な検索設定を提供する**．文書の平均長は 183.5 ページであり，LongDocURL [12]（85.6 ページ）の 2.1 倍以上である．



(a) Model overview of **CMDR-Embed** (b) **CMCL** (**C**ontextual **M**ultimodal **C**ontrastive **L**earning)

図 3 (a) CMDR-Embed の概要と，(b) 文脈の活用と同一文書内における識別性を向上させるための対照学習 CMCL.

3 提案モデル

図 3 に，CMDR-Embed のモデル概要と学習方法を示す．貢献は，(1) 各ページ自身の内容に加えて，文脈情報も捉えた埋め込みの構築が可能，(2) 文脈情報を活用しつつ，同一文書内における識別性を向上させるための新たな学習手法 CMCL の提案である．

3.1 モデル構造

文脈化マルチモーダル符号化 文脈情報を活用するため，複数ページ（チャンク）を同時に入力しエンコード後，各ページの埋め込みに分割する．まず，文書 $\mathbf{I} = \{I_1, \ldots, I_N\}$ が与えられたとき，窓幅 w，移動幅 s を用いて文書を複数のチャンクに分割する．t 番目のチャンク $\{I_{s(t-1)+1}, \ldots, I_{s(t-1)+w}\}$ は大規模視覚言語モデル（LVLM）に入力され，各ページのトークン表現 $\{\mathbf{E}_{s(t-1)+1}, \ldots, \mathbf{E}_{s(t-1)+w}\} \in \mathbb{R}^{w \times N^I \times D}$ に分割する．N^I は画像トークン数，D は埋め込み次元を表す．チャンクがオーバーラップされる（$s < w$）時，複数のチャンクに重複して含まれるトークンは，冗長な計算を避けるため，直前のチャンクにおける表現を採用する．最終的に，文脈化文書埋め込み系列 $\{\mathbf{E}^{\mathbf{I}}_1, \ldots, \mathbf{E}^{\mathbf{I}}_N\} \in \mathbb{R}^{N \times N^I \times D}$ を得る．

遅延相互作用 クエリと各ページとの類似度は，複数トークン埋め込み用いた検索手法である遅延相互作用（**LI**: Late Interaction）[16, 17] を用いる．クエリは，LVLM により符号化されたトークン埋め込み $\mathbf{E}^Q \in \mathbb{R}^{N^q \times D}$ として表現され，N^q はクエリトークン数を表す．クエリ q と k 番目のページ I_k との類似度スコアは，次式で計算される：

$$\mathbf{LI}(q, I_k) = \sum_{i \in [1, N^q]} \max_{j \in [1, N^I]} \langle \mathbf{E}^{q_i}, \mathbf{E}^{\mathbf{I}_j}_k \rangle. \tag{1}$$

3.2 文脈化マルチモーダル対照学習

文脈情報を考慮した埋め込みの最適化と，同一文書内における識別性の維持を両立するため，対照学習 CMCL を提案する．具体的には，InfoNCE [18] に基づく対照学習において，同一文書内の他チャンク，および同一チャンク内の別ページをハード負例（In-Doc Negatives $\mathcal{S}_{\text{doc}}$/ In-Seq Negatives $\mathcal{S}_{\text{seq}}$）として扱う．これにより，同一文書内の埋め込み間に対照制約を課すことで，文脈情報の過度な共有による表現崩壊を防ぎつつ，各チャンクが固有の情報を保持した識別的な表現を獲得できると期待される．

学習損失 CMCL は，$\mathcal{L} = \lambda\mathcal{L}_{\text{Context}} + (1-\lambda)\mathcal{L}_{\text{Batch}}$ を最小化するように学習する．正例，負例ページをそれぞれ I^+，I^- とすると，次式で計算される．

$$\begin{aligned}\mathcal{L}_{\text{Context}} = &-\log \frac{e^{\mathbf{LI}(q,I^+)/\tau}}{e^{\mathbf{LI}(q,I^+)/\tau} + \sum_{I_n^- \in \mathcal{S}_{\text{doc}}} e^{\mathbf{LI}(q,I_n^-)/\tau}} \\ &-\log \frac{e^{\mathbf{LI}(q,I^+)/\tau}}{e^{\mathbf{LI}(q,I^+)/\tau} + \sum_{I_n^- \in \mathcal{S}_{\text{seq}}} e^{\mathbf{LI}(q,I_n^-)/\tau}},\end{aligned} \tag{2}$$

$$\mathcal{L}_{\text{Batch}} = -\log \frac{e^{\mathbf{LI}(q,I^+)/\tau}}{e^{\mathbf{LI}(q,I^+)/\tau} + \sum_{I_n^- \in \mathcal{S}_{\text{batch}}} e^{\mathbf{LI}(q,I_n^-)/\tau}}, \tag{3}$$

τ は温度パラメータ，$\mathcal{S}_{\text{batch}}$ はバッチ内負例を表す．

4 実験

学習データ Common Crawal から収集した PDF を基に，Qwen2.5-VL 72B [21] を用いて，39,796 件の学習データを作成した．構築方法は付録に示す．

比較モデル 文脈情報を活用しない Non-Contextual Models を比較対象とする．具体的には，文書テキストをエンコードする 3 件の Text Retrievers，一般/文書画像ドメインで学習された画像ベースのモデル General Multimodal Retrievers/Multimodal Document Retrievers をそれぞれ 3 件，5 件採用した．

表 2 CMDR-Bench におけるページ検索結果．R: Recall，nD: nDCG．Overall は 2.2 節に示す各クエリカテゴリの平均値．

Model	Backbone	#Params	TC		CR		SU		MR		Overall	
			R@5	nD@5	R@5	nD@5	R@5	nD@5	R@5	nD@5	R@5	nD@5
Non-Contextual Models			*Text Retrievers*									
BM25 [1]	–	–	42.9	29.1	29.3	19.5	32.8	20.9	39.7	24.9	36.2	23.5
Contriever [2]	BERT-base	109M	53.6	35.7	23.8	15.0	36.8	23.5	40.5	26.4	38.6	25.1
NV-Embed-v2 [8]	Mistral-7B	7.9B	56.0	36.2	26.0	16.9	44.6	28.6	47.8	30.4	43.6	28.0
			General Multimodal Retrievers									
CLIP [19]	CLIP-large	428M	20.2	13.3	7.7	5.8	17.6	10.6	20.6	13.2	16.6	10.7
VLM2Vec [9]	Phi-3.5-V	4.2B	39.9	27.9	24.3	16.4	39.7	26.4	38.9	28.2	35.7	24.7
GME [20]	Qwen2-VL	2.2B	57.7	37.1	38.7	26.0	49.0	30.8	53.8	38.0	49.8	33.0
			Multimodal Document Retrievers									
VisRAG-Ret [6]	MiniCPM-V	3.4B	53.6	35.3	30.4	19.8	45.6	27.8	50.2	33.9	44.9	29.2
ColPali [5]	Paligemma	2.9B	59.5	38.8	37.6	24.2	48.5	31.0	55.1	35.7	50.2	32.4
+Finetune	Paligemma	2.9B	62.5	41.0	42.5	27.5	55.9	36.8	56.3	39.9	54.3	36.3
ColQwen [5]	Qwen2-VL	2.2B	59.5	37.7	39.2	27.2	51.0	30.3	54.7	35.0	51.1	32.6
+Finetune	Qwen2-VL	2.2B	66.1	45.8	50.8	34.8	57.8	38.7	60.3	42.2	58.8	40.4
Contextual Models (Ours)												
CMDR-Embed$_{\mathrm{Pali}}$ (Ours)	Paligemma	2.9B	67.3	45.9	44.8	30.2	71.6	50.6	67.6	51.4	62.8	44.5
CMDR-Embed$_{\mathrm{Qwen}}$ (Ours)	Qwen2-VL	2.2B	**85.1**	**66.0**	**54.7**	**40.3**	**81.9**	**64.8**	**78.5**	**59.9**	**75.1**	**57.7**

表 3 アブレーション評価．Overall スコアを報告する．

	R@5	nD@5
CMDR-Embed$_{\mathrm{Pali}}$	**62.8**	**44.5**
w/o In-Doc Negatives	59.2	41.6
w/o In-Seq Negatives	59.9	41.3
w/o CMCL (↪ Only In-Batch Negatives)	57.1	38.8
w/o LI (↪ Mean Pooled Single Vector)	38.7	25.2

4.1 評価結果と分析

文脈情報は性能向上に寄与するか？ 表 2 に示すように，文脈情報を活用する CMDR-Embed が，文脈を考慮しないモデルを大きく上回る性能を達成した．この性能向上は，学習データの性質によるものではない．同一のデータで学習した非文脈モデルの ColPali や ColQwen は，文脈モデル CMDR-Embed$_{\mathrm{Pali}}$ と CMDR-Embed$_{\mathrm{Qwen}}$ を下回る結果であった．

CMDR-Bench の特徴は何か？ 表 2 に示すように，画像検索モデルがテキスト検索モデルよりも高い性能を示し，さらに文脈情報を活用することで性能向上が確認できる．以上より，CMDR-Bench はマルチモーダル情報と文書文脈の双方を適切に評価できる挑戦的なベンチマークであることが分かる．

CMCL は性能に影響するか？ 表 3 に示すように，CMCL によって性能向上を確認した．さらに，同一文書内のチャンクを用いた二種類の負例はいずれも有効であることが分かった．

LI は性能に影響するか？ 表 3 より，LI は単一トークンによる類似度計算よりも，高い性能を示した．複数ページを同時に扱う提案手法においては，情報量を多く保持可能な LI が有効だと考えられる．

表 4 効率性分析．Overall スコアおよび，A100 GPU 一台による Index 化にかかる平均時間（page/ms）を報告する．

	R@5	nD@5	Index
ColPali + Finetune	54.3	36.3	37.3
CMDR-Embed$_{\mathrm{Pali}}$	62.8	44.5	27.7
CMDR-Embed$_{\mathrm{Pali}}$ w/o Overlap ($w = s$)	58.2	41.7	13.4

CMDR-Embed の効率性は？ 表 4 に示すように，CMDR-Embed$_{\mathrm{Pali}}$ はベースの ColPali と比較して高い性能を示すとともに，インデックス化時間も短く，高い効率性を示した．複数ページを同時に埋め込むことで forward 計算回数が減少するためである．

5 おわりに

文脈を考慮したマルチモーダル文書検索タスク CMDR を提案し，本分野における文脈情報活用の重要性を明らかにした．本研究は，産業上重要な実世界文書を対象とした情報検索の高度化に貢献する．

関連研究と議論 従来のマルチモーダル文書検索 [3, 4, 5, 6, 7] は，ページ単位の独立な一致に基づく設計により，文脈を考慮した検索が実現できていない．本研究では，文脈を考慮したマルチモダール検索タスクおよびベンチマーク CMDR/CMDR-Bench を提案し，文脈を統合的に表現する検索モデル CMDR-Embed を導入した．さらに，同一文書内での識別性を保つ対照学習 CMCL により，効率性を維持しつつ大幅な性能向上を達成した．

参考文献

[1] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. **Foundations and Trends® in Information Retrieval**, Vol. 3, No. 4, pp. 333–389, 2009.

[2] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. **arXiv:2112.09118**, 2021.

[3] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. SlideVQA: A dataset for document visual question answering on multiple images. In **AAAI**, pp. 13636–13645, 2023.

[4] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. **NeurIPS**, Vol. 37, pp. 95963–96010, 2024.

[5] Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. **arXiv:2407.01449**, 2024.

[6] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. **arXiv:2410.10594**, 2024.

[7] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In **CVPR**, pp. 24827–24837, 2025.

[8] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-Embed: Improved techniques for training llms as generalist embedding models. **arXiv:2405.17428**, 2024.

[9] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. **arXiv:2410.05160**, 2024.

[10] Max Conti, Manuel Faysse, Gautier Viaud, Antoine Bosselut, Céline Hudelot, and Pierre Colombo. Context is gold to find the gold passage: Evaluating and training contextual document embeddings. **arXiv:2505.24782**, 2025.

[11] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. **Pattern Recognition**, Vol. 144, p. 109834, 2023.

[12] Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In **ACL**, pp. 1135–1159, 2025.

[13] Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multi-modal retrieval for long documents. **arXiv:2501.08828**, 2025.

[14] Jian Chen, Ming Li, Jihyung Kil, Chenguang Wang, Tong Yu, Ryan Rossi, Tianyi Zhou, Changyou Chen, and Ruiyi Zhang. Visr-bench: An empirical study on visual retrieval-augmented generation for multilingual long document understanding. **arXiv:2508.07493**, 2025.

[15] Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. In **ICCV**, pp. 17443–17453, 2025.

[16] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In **SIGIR**, pp. 39–48, 2020.

[17] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. **arXiv:2402.03216**, 2024.

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. **arXiv:1807.03748**, 2018.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **ICML**, pp. 8748–8763, 2021.

[20] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In **CVPR**, pp. 9274–9285, 2025.

[21] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. **arXiv:2502.13923**, 2025.

[22] Zheng Liu, Ze Liu, Zhengyang Liang, Junjie Zhou, Shitao Xiao, Chao Gao, Chen Jason Zhang, and Defu Lian. Any information is just worth one single screenshot: Unifying search with visualized information retrieval. In **ACL**, pp. 19238–19261, 2025.

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. **arXiv:2106.09685**, 2021.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. **arXiv:1711.05101**, 2017.

[25] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with io-awareness. In **NeurIPS**, pp. 16344–16359, 2022.

[26] Ray Smith. An overview of the tesseract ocr engine. In **ICDAR**, pp. 629–633, 2007.

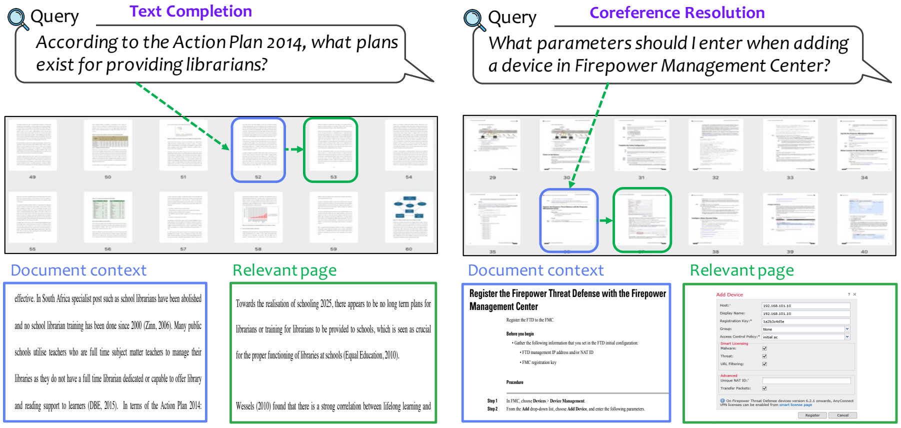


図 4　CMDR-Bench における Text Completion と Coreference Resolution の例.

表 5　学習データ構築における (1) で活用したプロンプト.

Prompt: Filter pages if requring contextual information

```
You are given an image of a single page from a multi-page PDF document. Determine whether the content of this page is
**self-contained** (can be fully understood on its own) or **not self-contained** (requires other pages).
Criteria for judgment:
- Self-contained (Yes):
  - The page presents complete information that can be understood without referring to other pages.
  - The text begins and ends naturally (not cut off in the middle).
  - Title pages or pages with no text are considered self-contained.
  - Figures, tables, or explanations are understandable on this page alone.
- Not self-contained (No):
  - The page refers to figures, tables, or sections on other pages.
  - The page is incomplete or depends on missing context.
  - The page starts in the middle of a sentence, paragraph, or table.
Output: [Yes/No]
```

表 6　学習データ構築における (3) で活用したプロンプト.

Prompt: Generate queries requiring contextual information

```
Create one short retrieval-oriented query in English following the conditions below.
1. The query should appear to be about Image 1, but the correct answer is found only in Image 2.
- At first glance, it should seem that the answer could be in Image 1.
- Only by reading both pages carefully should it become clear that the answer is actually in Image 2.
2. The query must require contextual understanding across both pages, not a simple keyword search.
3. Query Type: {query type}
4. Do not quote or copy any text from Image 2 directly into the query.
- Do not include explicit hints or keywords that would guide the reader to look into Image 2.
5. Only output the query in the specified format without any additional explanation.
6. Use only one question word and keep the query in a single sentence.
Output: <Your query here>
```

A　付録

クエリ作成　図 4 に，CMDR-Bench における Text Completion と Coreference Resolution の例を示す．

学習データ構築　Common Crawl から収集した PDF に対し，クエリ付与を以下の 3 段階で行った．(1) Qwen2.5VL 72B [21] を用いて，各ページについて内容が文脈情報を必要とするか（self-contained か否か）を分類する．プロンプトを表 5 に示す．(2) 文脈が必要と判定されたページを入力とし，UniSE [22] を用いて文脈情報となページを検索し，（文脈情報ページ，関連ページ）のページペアを構築する．プロンプトには，”*Retrieve the page that provides important contextual information for better understanding the given page.*”を使用した．(3) 構築したページペアを Qwen2.5VL 72B に入力し，文脈情報を必要とするクエリを生成する．プロンプトを表 6 に示す．

実装の詳細　窓幅 w を 4，移動幅を 2 とし，CMCL の損失に関する λ を 0.1 とした．LLM には LoRA [23] を適用し，他のパラメータを固定したまま学習した．8 枚の A100-80G GPU 上で 3 エポックの訓練を行い，最適化には AdamW [24]，さらに，FlashAttention [25] を使用した．バッチサイズは 96 とし，学習率を 2e-4 とした．また，温度パラメータ τ を 0.02 に設定した．OCR テキスト抽出には Tesseract [26] を使用した．