VDocRAG: 視覚的文書に対する検索拡張生成

田中涼太 [1,2]　壹岐太一 [1]　長谷川拓 [1]　西田京介 [1]　齋藤邦子 [1]　鈴木潤 [2]
[1] 日本電信電話株式会社　NTT 人間情報研究所
[2] 東北大学
{ryota.tanaka, taichi.iki, taku.hasegawa, kyosuke.nishida, kuniko.saito}@ntt.com,
jun.suzuki@tohoku.ac.jp

概要

視覚的に表現された文書から成るコーパスを知識源に持つ新たな検索拡張生成（RAG）フレームワークである VDocRAG を提案する．VDocRAG は多様な文書を画像形式で統一的に理解することで，視覚的文書に含まれる図や表などの視覚情報を直接利用できる．VDocRAG の性能向上を目的として，大規模視覚言語モデルを検索タスクに適応させる新たな自己教師あり事前学習タスクを提案する．更に，多様な文書形式を網羅するオープンドメイン視覚文書質問応答データセットである OpenDocVQA を導入する．実験により，VDocRAG は従来のテキストベース RAG を大幅に上回る性能を示し，優れた汎化能力を有することが確認された．

1　はじめに

大規模言語モデル（LLM）は様々な自然言語処理タスクにおいて優れた性能を示している [1, 2]．一方で，事実に反した出力をする問題を抱えている [3, 4]．この問題を解決するために，検索拡張生成（**RAG**: Retrieval-Augmented Generation）が提案されている [5, 6] が，従来の RAG は，検索の知識源となる文書がテキストのみで記述されていることを仮定している．一方，実世界の情報の多くは，図表などの視覚情報を含む視覚的文書に格納されている．

現実世界に存在する多様な文書の理解を目指して，視覚的文書を対象とした質問応答タスクである DocumentVQA [7, 8, 9, 10] が活発に研究されている．ここで，従来の DocumentVQA では，質問の対象となる文書が事前に与えられているため，多くの質問は検索を必要としない設定となっている．この制約により，任意の文書を知識源として検索しつつ回答を行う，オープンドメインな DocumentVQA モデルの実現には至っていない．

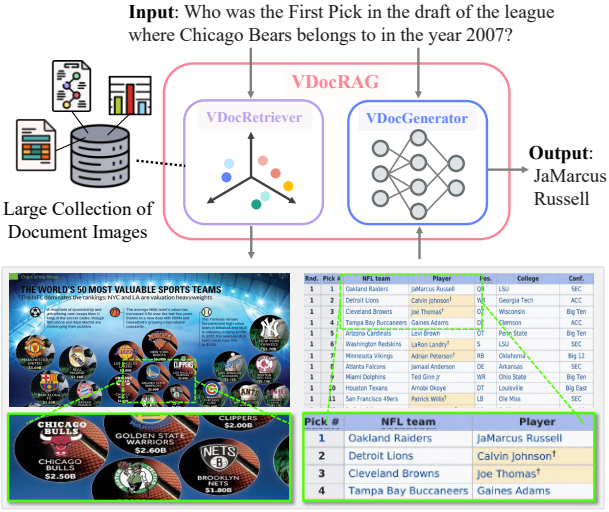


図 1　VDocRAG の概要と OpenDocVQA の例．

本研究では，大規模視覚言語モデル（LVLM）[11, 12] を活用した新たな RAG フレームワーク **VDocRAG** を提案する．VDocRAG は，多様な文書を画像形式で統一的に理解することで，視覚的文書に含まれる図や表などの視覚情報を直接利用できる．図 1 に示すように，VDocRAG は質問に関連する文書画像を検索する VDocRetriever と，検索された文書画像を用いて回答を生成する VDocGenerator で構成されている．VDocRAG の性能向上を目指して，LVLM の高い画像理解・テキスト生成能力を活用し，文書内のテキストを画像埋め込み表現に圧縮する事前学習タスクである **R**epresentation **C**ompression via **R**etrieval and **G**eneration（**RCR**，**RCG**）を提案する．

さらに，多様な文書形式を網羅する初のオープンドメイン DocumentVQA データセット **OpenDocVQA** を提案する．OpenDocVQA は，視覚的文書を対象とした検索と QA モデルの学習・評価を行うための包括的なリソースを提供する．実験の結果，VDocRAG は従来のテキストベース RAG を大幅に上回る性能を示し，高い汎化性能を有することを確認した．

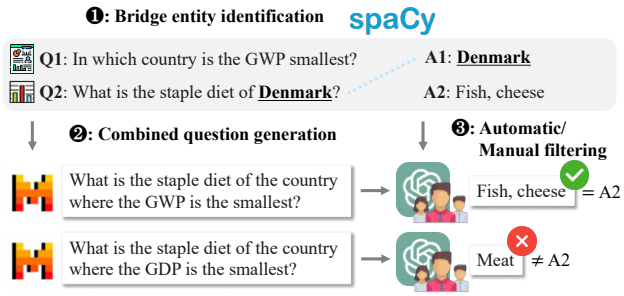


図 2　マルチホップ質問の作成方法.

2　OpenDocVQA

2.1　問題定義

N 枚の文書画像で構成されるコーパス $\mathcal{I} = \{I_1, ..., I_N\}$ と質問文 Q が与えられ，関連する k 枚の画像 $\hat{\mathcal{I}} \in \mathcal{I}$（$k \ll N$）を特定し，回答 A を出力する．本タスクは，以下の二つのステップに分解できる．**視覚的文書検索**: Q および $\mathcal{I}$ を与え，モデルは回答を導出するために必要な k 枚の画像 $\hat{\mathcal{I}}$ を検索する．**DocumentVQA**: Q と検索で得られた画像集合 $\hat{\mathcal{I}}$ を入力とし，モデルは回答 A を生成する．

OpenDocVQA は多様な文書形式を網羅する複数のデータセットから成る．各データセットが提供する特定の文書形式から検索する **Single-pool** と，データセットを横断して検索する **All-pool** で評価する．

2.2　データ収集

DocumentVQA データ収集　既存の 7 件の DocumentVQA データセット [7, 8, 13, 9, 10, 14, 15] を収集し，フィルタリングを行った．既存データセットの質問文の多くは，文書を参照しなければ回答できない文脈依存性（例: *What is the page number?*）がある．OpenDocVQA タスクに適用するために，文脈依存性を持つ質問文を削除した．具体的には，ヒューリスティックな基準（付録参照）を適用後，全サンプルが文脈非依存となるように人手で精査した．

TableQA の改変　Wikipedia の表を検索し，回答を行う TableQA データセット [16] を改変した．具体的には，元データセットは表がテキスト形式で提供されているため，対応する表のスクリーンショット画像を Wikipedia から取得し，知識源とした．

マルチホップ質問作成　複数の文書を参照するマルチホップ推論の能力を向上させるために，収集された QA ペアを用いて，以下の手順で新たなデータセット MHDocVQA を作成した（図 2 参照）．1）

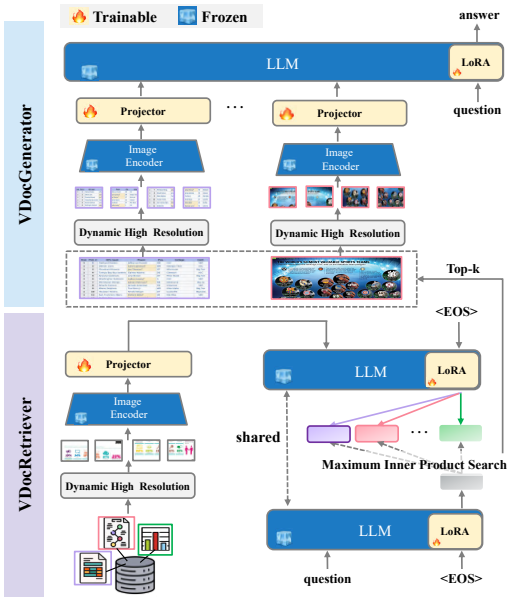


図 3　VDocRAG の概要．LLM，画像エンコーダ，プロジェクタから成る LVLM を用いて，文書を画像として検索し，質問に回答する．

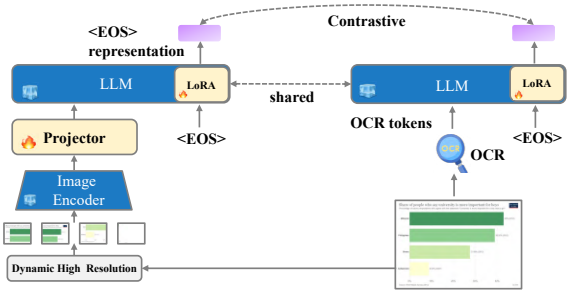
spaCy [17] を使用し，シングルホップ質問の回答に含まれるエンティティ（例: *Denmark*）を特定し，このエンティティが含まれる別のシングルホップ質問を発見した．2）Mixtral-8x22B [2] を用いて 2 つの質問文を結合し，マルチホップ質問を作成した．3）GPT-4o [1] を使用し，2 つのシングルホップ質問およびその回答に基づいてマルチホップ質問に回答し，予測された回答がシングルホップ質問の回答と一致しない場合，削除した．最後に，フィルタリング後の質問を人手で精査し，品質を確保した．

負例画像マイニング　まず，大規模画像データセット COYO-700M [18] を用いて，画像中の OCR テキストを抽出した．次に，OCR テキストが質問文と最も高い単語一致率であり，かつ，OCR テキストに回答を含まない画像を負例として収集した．

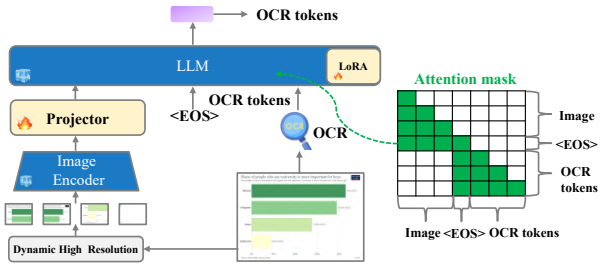
統計情報　OpenDocVQA は，新規で構築した MHDocVQA を含む 9 件のソースデータセットから構成されており，230,858 枚の文書画像をコーパスとして持ち，43,474 件の QA ペアが含まれるオープンドメイン DocumentVQA データセットである．従来データセットとの比較は，付録で議論する．

3　提案モデル

図 3 に示すように，VDocRAG は VDocRetriever と VDocGenerator の 2 つのモジュールから構成される．文書を全て画像形式に統一することで，単一のモデルで多様な形式の文書理解を実現する．



(a) Representation Compression via Retrieval (RCR)



(b) Representation Compression via Generation (RCG)

図 4　我々が提案する自己教師あり事前学習タスク．

3.1　モデル構造

高解像度画像エンコーディング　動的クロッピング [19] を適用し，アスペクト比を維持し画像を複数のパッチに分割する．各パッチを画像エンコーダと 2 層の MLP を通して視覚的文書特徴 $\mathbf{z}_\mathrm{d}$ に変換する．

VDocRetriever　LVLM により，質問文と文書画像を独立してエンコードする．まず，質問文と $\mathbf{z}_\mathrm{d}$ の終端に<EOS>を付加する．次に，それらを LLM に入力し，最終層の<EOS>表現を用いて，質問文および視覚的文書の埋め込み（$\mathbf{h}_\mathrm{q}, \mathbf{h}_\mathrm{d}$）を獲得する．最後に，質問文に対する類似度スコア $\mathrm{SIM}(\mathbf{h}_\mathrm{q}, \mathbf{h}_\mathrm{d}) = \frac{\mathbf{h}_\mathrm{q}^\top \mathbf{h}_\mathrm{d}}{\|\mathbf{h}_\mathrm{q}\| \|\mathbf{h}_\mathrm{d}\|}$ が高い k 件の画像 $\hat{\mathcal{I}}$ を検索する．

VDocGenerator　LVLM を用いて，VDocRetriever によって取得された k 件の文書画像 $\hat{\mathcal{I}}$ と質問文 Q を基に，回答 A を生成する．検索結果と質問文を連結し，LLM に与えることで回答を生成する．

3.2　自己教師あり事前学習

図 4 に示すように，LVLM を視覚的文書検索へ適用するために，<EOS>トークンに画像表現を圧縮するための新たな事前学習タスクを提案する．事前学習の目的関数は，$\mathcal{L} = \mathcal{L}_\mathrm{RCR} + \mathcal{L}_\mathrm{RCG}$ で定義される．

検索による表現圧縮（RCR）　LVLM の画像理解能力を活用し，OCR テキストに対応する画像を検索する対照学習タスクを通じて画像表現を圧縮する．図 4a に示すように，まず，正例の OCR テキスト-画像ペア $(\mathbf{h}_\mathrm{o}, \mathbf{h}_{\mathrm{d}^+})$ を構築する．次に，バッチ内負例サンプリングを用いて，対照損失を計算する:

$$\mathcal{L}_\mathrm{RCR} = -\log \frac{\exp(\mathrm{SIM}(\mathbf{h}_\mathrm{o}, \mathbf{h}_{\mathrm{d}^+})/\tau)}{\sum_{i \in \mathcal{B}} \exp(\mathrm{SIM}(\mathbf{h}_\mathrm{o}, \mathbf{h}_{\mathrm{d}_i})/\tau)}, \tag{1}$$

τ は温度パラメータ，$\mathcal{B}$ はバッチサイズを表す．

生成による表現圧縮（RCG）　アテンションマスクを工夫することで，LVLM のテキスト生成能力を活用する表現学習を提案する．図 4b に示すように，<EOS>を含む画像トークン表現は，自己回帰によって得られる．一方，後続の L 個の OCR トークンでは，画像トークンをマスクし，<EOS>と過去の OCR トークンに対してのみアテンションを許可する．これより，画像表現を<EOS>に集約することができる．損失関数は以下のように定義される:

$$\mathcal{L}_\mathrm{RCG} = -\frac{1}{L} \sum_{i=1}^{L} \log p(y_i | y_{<i}, \texttt{<EOS>}), \tag{2}$$

y_i は i 番目の OCR トークンを表す．

3.3　2 段階ファインチューニング

まず，正例の質問文-文書ペアとバッチ内の負例を用いた対照学習により，VDocRetriever をファインチューニングする．次に，学習済み VDocRetriever を用いて，関連する上位 k 件の文書を $\mathcal{I}$ から検索する．最後に，質問文と検索結果を入力として，VDocGenerator を次単語予測損失により訓練する．

4　実験

データセット　表 3 の通り，zero-shot 設定（ChartQA と SlideVQA）と教師あり設定（InfoVQA と DUDE）で評価した．事前学習には，DocStruct4M [27] から抽出した 50 万ペアを使用した．

実装　VDocRAG の初期化には，Phi3V [11] を使用した．VDocRetriever で取得した上位 3 件の文書を利用した．詳細は，付録に示す．

ベースライン　検索モデルには，OCR テキストをエンコードする 6 モデルと画像ベースの 3 モデルを採用した（表 1）．事前学習は行わず，OpenDocVQA のみで学習した提案モデル（VDocRetriever–）と Phi3 を主な比較対象とした．QA モデルには，OCR テキストを入力とする Phi3 を検索と回答生成に用いる TextRAG と，検索を行わない Phi3 を採用した．

評価指標　検索性能の評価には，**nDCG@5** を用いた [28, 29]．InfoVQA と DUDE には **ANLS** [30]，ChartQA には **Relaxed Accuracy** [13]，SlideVQA には **F1** を評価指標として使用した．

表 1 OpenDocVQA の結果．Single-pool/All-pool 設定の結果．文書を OCR テキスト，または，画像で表現する．GT は正解文書のみをモデルに与える．緑色の結果は，対応する学習データを使用した場合の結果．

(a) 視覚的文書検索の結果．

Model	ChartQA	SlideVQA	InfoVQA	DUDE
Off-the-shelf				
BM25 [20]	54.8/15.6	40.7/38.7	50.2/31.3	57.2/47.5
Contriever [21]	66.9/59.3	50.8/46.5	42.5/21.0	40.6/29.7
E5 [22]	74.9/66.3	53.6/49.6	49.2/26.9	45.0/38.9
GTE [23]	72.8/64.7	55.4/49.1	51.3/32.5	42.4/36.0
E5-Mistral [24]	72.3/70.0	63.8/57.6	60.3/33.9	52.2/45.2
CLIP [25]	54.6/38.6	38.1/29.7	45.3/20.6	23.2/17.6
DSE [26]	72.7/68.5	73.0/67.2	67.4/49.5	55.5/47.6
Trained on OpenDocVQA				
Phi3 [11]	72.5/65.3	53.3/48.4	53.2/33.0	40.5/32.0
VDocRetriever–	84.2/74.8	71.0/64.7	66.8/52.6	48.4/40.6
VDocRetriever	**86.0/76.3**	**77.3/73.0**	**72.9/55.2**	**57.7/50.6**

(b) DocumentVQA の結果．

Model	ChartQA	SlideVQA	InfoVQA	DUDE
Phi3	20.0/20.0	20.3/20.3	34.9/34.9	23.1/23.1
TextRAG	28.0/28.0	28.6/28.0	40.5/39.1	40.1/35.7
$\text{TextRAG}_{\text{GT}}$	36.6/36.6	27.8/27.8	45.6/45.6	55.9/55.9
VDocRAG	**52.0/48.0**	**44.2/42.0**	**56.2/49.2**	**48.5/44.0**
$\text{VDocRAG}_{\text{GT}}$	74.0/74.0	56.4/56.4	64.6/64.6	66.4/66.4

表 2 Single-pool 設定の検索タスクにおける ablation 評価．

Model	SlideVQA	InfoVQA
VDocRetriever	**77.3**	**72.9**
w/o RCR	$75.9_{-1.4}$	$71.1_{-1.8}$
w/o RCG	$71.7_{-5.6}$	$68.8_{-4.1}$
w/o RCG & RCR	$71.0_{-6.3}$	$66.8_{-6.1}$
w/o LLM & Projector (↪ CLIP)	$43.7_{-33.6}$	$37.9_{-35.0}$

4.1 評価結果と分析

検索に関する結果 表 1a の通り，VDocRetriever– は同一条件下において，Phi3 よりも大幅に高い性能を達成した．更に，VDocRetriever は未知のデータセット（ChartQA, SlideVQA）において，従来の検索モデルを上回る zero-shot 性能を示した．特に，DSE は提案モデルと同じ LVLM を初期化に使用し，13.7 倍のデータ量でファインチューニングされているにも関わらず，提案モデルが優位性を示した．これは，新たな事前学習とデータセットが，従来では対応が困難な要素をカバーすることを示唆する．

RAG に関する結果 表 1b の通り，VDocRAG は Phi3 や TextRAG を上回る性能を達成した．また，正解文書の付与により，大幅な性能向上を示した．これは，検索の性能改善に余地があること，検索ノイズに対する頑健性が課題であることを示唆する．

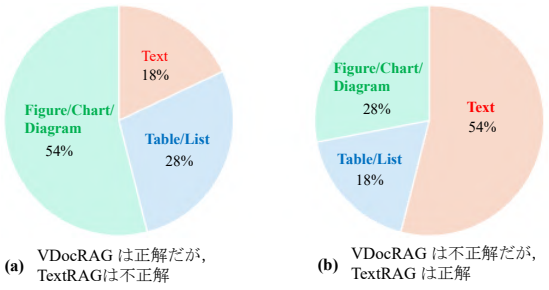


図 5 生成結果の要因調査．

事前学習は性能向上に寄与するか？ 表 2 の通り，提案モデルは事前学習を行わないモデルよりも優れた結果を達成した．それぞれの事前学習タスクを除去した場合，性能が低下したことから，RCR と RCG が補完的に寄与することが示された．

LLM は視覚的文書検索を促進するか？ 表 2 の通り，LLM を削除し，CLIP テキスト/画像エンコーダを使用した場合，検索性能が大幅に低下した．これは，LLM が詳細な視覚情報を捉え，意味理解を向上させる役割を担うことを示唆している．

人手評価 VDocRAG と TextRAG の生成結果に対して，50 件の正解例と 50 件の誤答例を対象に，生成結果の要因を人手で分析した．図 5a の通り，VDocRAG は特に視覚データ理解を得意とする．一方，VDocRAG は主に OCR 性能の影響により，テキスト主体の文書に苦戦する（図 5b 参照）．また，TextRAG は質問文と高いテキストの重複を含む文書において，正答を出力する傾向が見られた．

5 おわりに

多様な現実世界の文書を理解可能な VDocRAG を提案し，従来の TextRAG を大幅に上回る性能を示した．これにより，視覚表現された文書を対象とした RAG の開発に向けた新たな方向性を示した．

関連研究と議論 同時期の研究に文書画像に対する検索 [26, 29] や RAG [31] がある．これらは，特定の文書形式を対象にしたデータセットを使用しており，さらに検索に特化した学習なしで LVLM を活用しているため，多様な文書を知識源とする RAG の実現が困難であった．本研究では，1) 初のオープンドメイン DocumentVQA データセット「OpenDocVQA」の導入，2) LVLM を活用し，画像表現を圧縮する検索のための事前学習タスクの提案，を行った．

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. **arXiv:2303.08774**, 2023.

[2] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. **arXiv:2401.04088**, 2024.

[3] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **ACL**, pp. 9802–9822, 2023.

[4] Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **NAACL**, pp. 5506–5521, 2024.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **NIPS**, 2020.

[6] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In **ICML**, pp. 3929–3938, 2020.

[7] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In **WACV**, pp. 2200–2209, 2021.

[8] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In **AAAI**, pp. 13878–13888, 2021.

[9] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In **WACV**, pp. 1697–1706, 2022.

[10] 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.

[11] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. **arXiv:2404.14219**, 2024.

[12] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. **arXiv:2408.12637**, 2024.

[13] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In **Findings of ACL**, pp. 2263–2279, 2022.

[14] Jordy Landeghem, Rubén Tito, Lukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In **ICCV**, pp. 19528–19540, 2023.

[15] Liang Zhang, Anwen Hu, Jing Zhang, Shuo Hu, and Qin Jin. MPMQA: multimodal question answering on product manuals. In **AAAI**, pp. 13958–13966, 2023.

[16] Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo, and Edward Choi. Open-WikiTable : Dataset for open domain question answering with complex reasoning over table. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of ACL**, pp. 8285–8297, 2023.

[17] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[18] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

[19] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **EMNLP Findings**, pp. 2841–2858, 2023.

[20] 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.

[21] 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.

[22] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. **arXiv:2212.03533**, 2022.

[23] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. **arXiv:2308.03281**, 2023.

[24] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **ACL**, pp. 11897–11916, 2024.

[25] 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.

[26] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. **arXiv:2406.11251**, 2024.

[27] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. **arXiv:2403.12895**, 2024.

[28] Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. Resources for brewing beir: Reproducible reference models and an official leaderboard, 2023.

[29] 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.

[30] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In **ICCV**, pp. 4290–4300, 2019.

[31] 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.

[32] Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. $\text{DuReader}_{\text{vis}}$: A Chinese dataset for open-domain document visual question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of ACL**, pp. 1338–1351, 2022.

[33] 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.

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

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

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

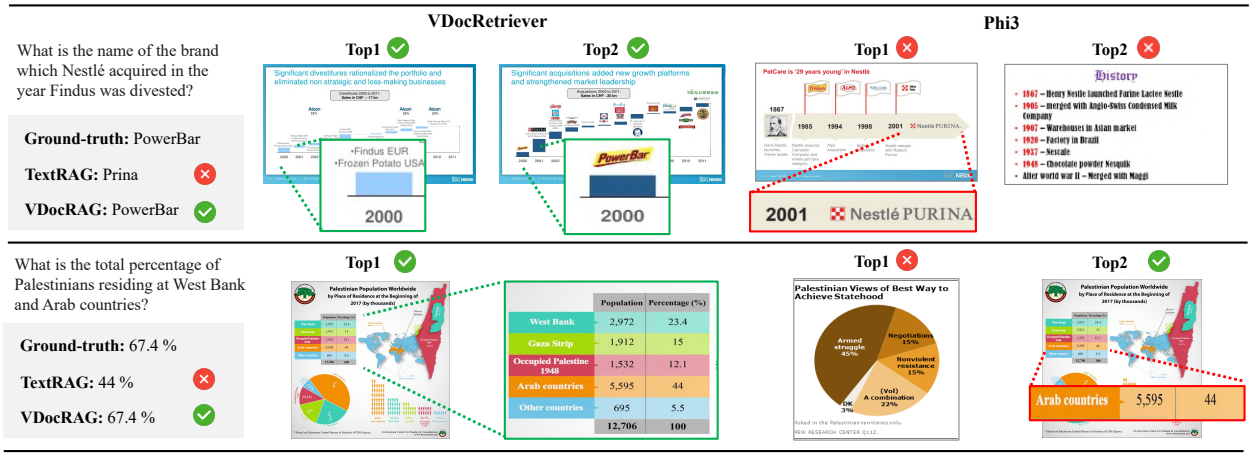


図 6　VDocRAG と TextRAG の生成結果に関する比較.

A　付録

表 3　OpenDocVQA に含まれるデータセット．† はマルチホップ推論を表す．Filtered はフィルタリング後に残ったサンプルの割合を示す.

Dataset	Document Type	%Filtered	#Images	#Train&Dev	#Test
DocVQA	Industry	84.8	12,767	6,382	–
InfoVQA	Infographic	61.2	5,485	9,592	1,048
VisualMRC	Webpage	71.9	10,298	6,126	–
ChartQA	Chart	94.0	20,882	–	150
OpenWikiTable	Table	100.0	1,257	4,261	–
DUDE	Open	92.3	27,955	2,135	496
MPMQA	Manual	81.7	10,018	3,054	–
SlideVQA†	Slide	66.7	57,617	–	760
MHDocVQA†	Open	100.0	28,550	9,470	–

表 4　データセット比較．文書には **T**able, **L**ist, **F**igure, **C**hart, **D**iagram が含まれる．回答形式は抽出型 (Ext) と要約型 (Abs) に大別できる.

	ViDoRe	Dureader$_{vis}$	OpenDocVQA
Retrieval	✓	✓	✓
QA	✗	✓	✓
Context-Independent	✗	✓	✓
Visual Semantic Search	✓	✗	✓
Multi-Hop	✗	✗	✓
Document Contents	T, L, F, C, D	T, L	T, L, F, C, D
Answer Types	–	Ext	Ext, Abs
#Document Types	6	1	Open
#QAs	3,810	15,000	43,474
#Images (Pages)	8,310	158,000	230,858

DocumentVQA のフィルタリング基準　以下の基準のいずれかに該当するサンプル全てをデータセットから削除した．基準を適用後，人手で精査を行った.

- 質問に「this」,「these」,「those」を含む 1 つ以上の指示代名詞が含まれている
- 質問に「she」,「he」,「her」,「his」,「him」を含む 1 つ以上の人称代名詞が含まれている
- 質問に「the document」や「mention」を含む特定のキーワードが含まれている
- 質問に数字以外のエンティティが含まれていない
- 質問が 6 単語未満である

表 5　Single-pool 設定における ablation 評価.

Model	Retrieval		QA	
	SlideVQA	InfoVQA	SlideVQA	InfoVQA
VDocRAG	**77.3**	**72.9**	**44.2**	**56.2**
w/o MHDocVQA	75.0$_{-2.3}$	71.4$_{-1.5}$	43.4$_{-0.8}$	53.8$_{-2.4}$
w/o except MHDocVQA	68.8$_{-8.5}$	61.7$_{-11.2}$	41.1$_{-3.1}$	44.0$_{-12.2}$

従来データセットとの比較　表 4 に従来データセット（ViDoRe [29] と Dureader$_{vis}$ [32]）を比較を示す．OpenDocVQA は主に 3 つ特長を持つ．1）OpenDocVQA は，オープンドメインな文書形式に対応する初の大規模 DocumentVQA データセットである．一方，ViDoRe は 6 種類の文書形式に関する検索タスクのみに対応しており，Dureader$_{vis}$ の文書は Web ページに限定している．2）OpenDocVQA の質問は文脈非依存であり，視覚情報を活用した検索を必要とする．一方，ViDoRe の質問は文脈依存であり，Dureader$_{vis}$ は BM25 などの語彙ベース手法でも十分な性能が得られる．3）ViDoRe や Dureader$_{vis}$ とは異なり，OpenDocVQA では抽出型（例: スパン，リスト）や生成型（例: 計算，カウント）の回答形式を伴うマルチホップ推論が必要となり，より挑戦的な設定である.

実装の詳細　LLM には LoRA [33] を適用し，他のパラメータを固定したまま Projector のみを更新する．8 枚の A100-80G GPU 上で 1 エポックの訓練を行い，最適化には AdamW [34]，さらに，FlashAttention [35] を使用した．バッチサイズは，事前学習時に 16，ファインチューニング時に 64 とした．また，温度パラメータ τ を 0.01 に設定した．OCR テキスト抽出には Tesseract [36] を使用した.

OpenDocVQA は性能向上に寄与するか？　表 5 の通り，MHDocVQA を除いた場合，性能が低下する．これは，MHDocVQA が OpenDocVQA 内の他のデータセットとは異なる推論能力を必要とすることを示唆する．更に，MHDocVQA を除く全てのデータセットを削除すると，性能が大幅に低下した．これは，我々が収集したデータセットが LVLM の文書検索・理解において不足する能力を効果的に補完可能であることを示す.

出力例　図 6 の通り，上部の例では，VDocRAG が複数スライドに渡るグラフ理解とマルチホップ推論を必要とする質問に対して，正解を出力している．下部の例では，VDocRAG が複数の行や列に跨る表の解析を必要とする質問に対しても優れた性能を示している．一方，TextRAG はテキスト情報のみに依存しているため，テキストの表面的な理解に留まり，不正確な予測になっている.