CVPR 2025

VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents

Ryota Tanaka[1,2] Taichi Iki[1] Taku Hasegawa[1] Kyosuke Nishida[1] Kuniko Saito[1] Jun Suzuki[2]
[1]NTT Human Informatics Laboratories, NTT Corporation [2]Tohoku University
https://vdocrag.github.io

- Problem / objective
  - Retrieval-Augmented Generation (RAG)

- Contribution / Key idea
  - **VdocRAG**: A new RAG framework, which directly understand diverse real-world documents purely from visual features
    - Self-supervised pre-training tasks, designed for document retrieval-oriented adaptation of LVLMs, by compressing visual document representations
    - OpenDocVQA: the first unified open-domain Document VQA dataset with diverse documents

전유진

VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents

CVPR 2025

OpenDocVQA Task and Dataset - Task

- OpenDocVQA
  a. N document images, question Q가 주어졌을때, question과 관련 있는 k개의 images 를 찾아 답변하기.
  b. Visual document retrieval + DocumentVQA

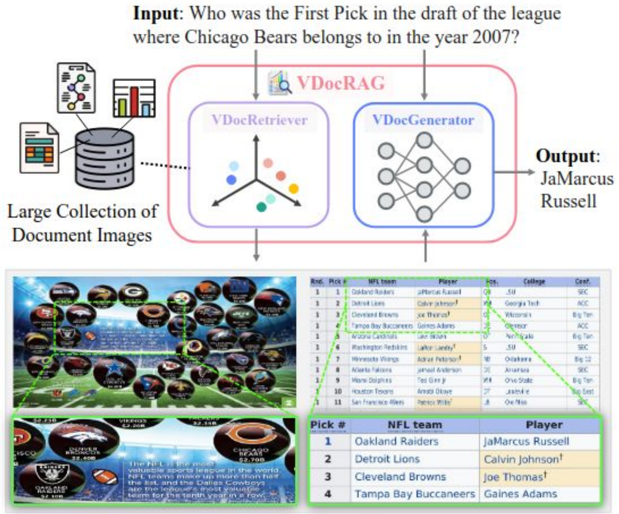


Figure 1. Our framework of VDocRAG and examples from Open-DocVQA. VDocRAG consists of VDocRetirver and VDocGenerator, which can retrieve relevant documents and generate answers by understanding the original appearance of documents.

	Input	Output
Visual document retrieval	N document images $\mathcal{I} = \{I_1, ..., I_N\}$ A question Q	Relevant k images $\hat{\mathcal{I}} \in \mathcal{I}$, where $k \ll N$
DocumentVQA	Relevant k images $\hat{\mathcal{I}} \in \mathcal{I}$, where $k \ll N$ A question Q	Answer A

전유진

OpenDocVQA Task and Dataset - Dataset

- Filtering of DocumentVQA datasets
- Reformulation of TableQA dataset
- Creation of new multi-hop questions
- Negative candidates mining

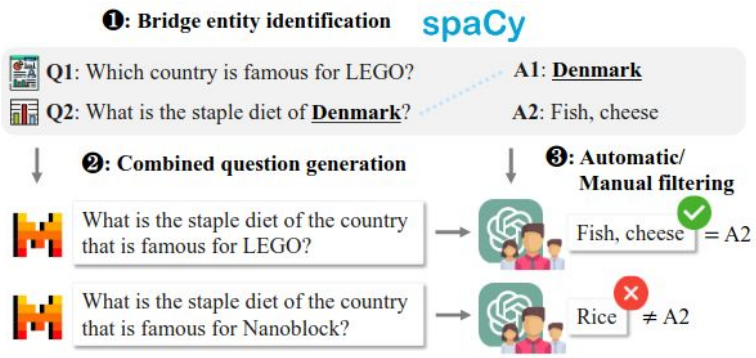


Figure 2. Process of creating multi-hop DocumentVQA questions.

	ViDoRe [17]	Dureader$_{vis}$ [46]	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	206,267

Table 1. Comparison of related datasets. Document contents include (T)able, (L)ist, (F)igure, (C)hart, and (D)iagram. Answer types are Extractive (Ext) and Abstractive (Abs).

- # Overview

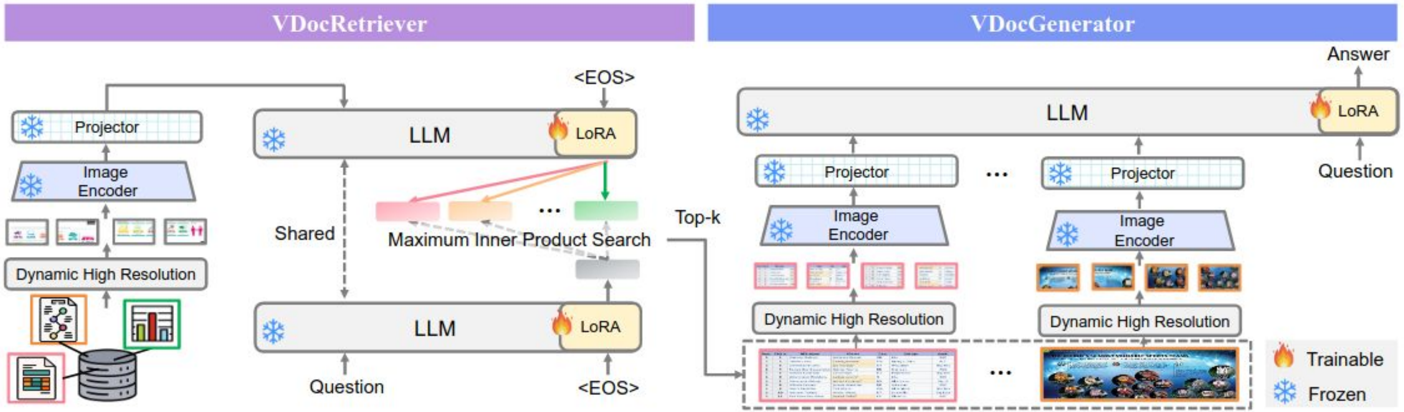


Figure 3. Overview of our VDocRAG model. VDocRetriever retrieves document images related to the question from a corpus of document images, and VDocGenerator uses these retrieved images to generate the answer.

- **Architecture Overview**
  - Dynamic high-resolution image encoding
    - Input: Document image, Output: Visual document features $\mathbf{z}_d$
    - 과정: 이미지 크롭(336x336)해서 encoding하고 2-layer MLP 통해 projection
  - VDocRetriever
    - LVLM-based dual-encoder architecture: queries와 document images를 독립적으로 인코딩
      - i. Question + <EOS> token --LLM-> question embeddings $\mathbf{h}_q$
      - ii. Visual document features + <EOS> token --LLM-> visual document embeddings $\mathbf{h}_d$
    - Maximum inner product search를 통해 유사도 높은 top-k documents 검색 $\mathrm{SIM}(\mathbf{h}_q, \mathbf{h}_d) = \frac{\mathbf{h}_q^\top \mathbf{h}_d}{\|\mathbf{h}_q\| \|\mathbf{h}_d\|}$
  - VDocGenerator
    - LLM input: Retrieved k documents 인코딩 결과 + question

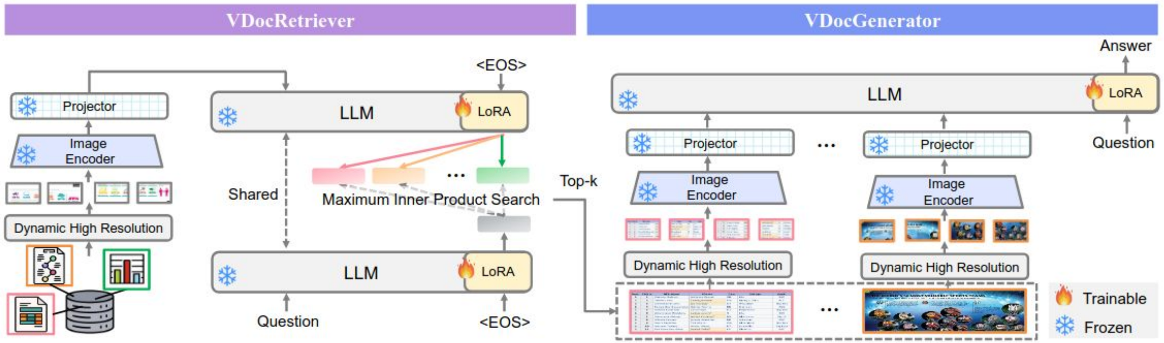


Figure 3. Overview of our VDocRAG model. VDocRetriever retrieves document images related to the question from a corpus of document images, and VDocGenerator uses these retrieved images to generate the answer.

CVPR 2025

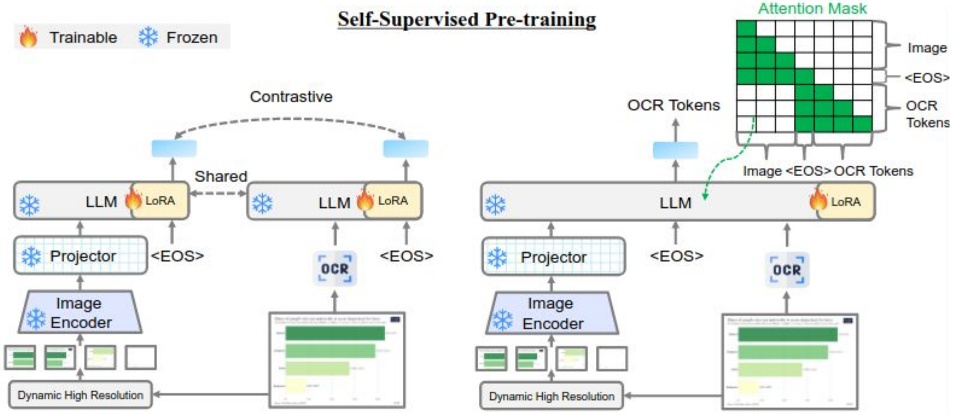
● Self-Supervised Pre-training Tasks

- ❑ 목표: To transfer the powerful abilities of LVLMs to facilitate their usage in visual document retrieval
- ❑ 그래서, entire image representation을 <EOS> token에 compress하기 위한 2가지 self-supervised pretraining tasks를 제안.
  a. Document image에서 추출한 OCR text을 psuedo target으로 사용.
  b. Full pre-training objectives: $\mathcal{L} = \mathcal{L}_{\text{RCR}} + \mathcal{L}_{\text{RCG}}$
  c. Representation Compression via Retrieval (RCR)
     i. Contrastive learning, for document-OCR text pairs (InfoNCE Loss)

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

  d. Representation Compression via Generation (RCG)
     i. Representation learning

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

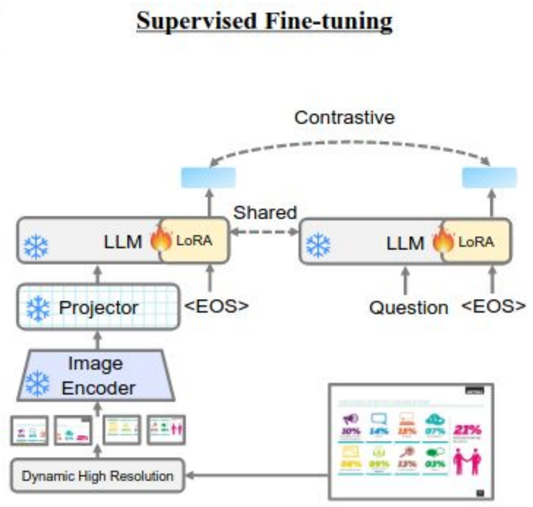


(a) Representation Compression via Retrieval (RCR) (b) Representation Compression via Generation (RCG)

Figure 4. Our pre-training tasks using unlabeled documents and fine-tuning in VDocRetriever.

전유진

- **Supervised Fine-tuning**
  - ❏ VDocRetriever
    a. Contrastive learning, for document-query pairs (InfoNCE Loss)
  - ❏ VDocGenerator
    a. Next-token prediction objective



(c) Visual Document Retrieval

- **Experiments**

Model	Init	Docs	Scale	#PT	#FT	ChartQA Single	ChartQA All	SlideVQA Single	SlideVQA All	InfoVQA Single	InfoVQA All	DUDE Single	DUDE All
Off-the-shelf													
BM25 [52]	–	Text	0	0	0	54.8	15.6	40.7	38.7	50.2	31.3	57.2	47.5
Contriever [22]	BERT [12]	Text	110M	1B	500K	66.9	59.3	50.8	46.5	42.5	21.0	40.6	29.7
E5 [59]	BERT [12]	Text	110M	270M	1M	74.9	66.3	53.6	49.6	49.2	26.9	45.0	38.9
GTE [34]	BERT [12]	Text	110M	788M	3M	72.8	64.7	55.4	49.1	51.3	32.5	42.4	36.0
E5-Mistral [60]	Mistral [23]	Text	7.1B	0	1.85M	72.3	70.0	63.8	57.6	60.3	33.9	52.2	45.2
NV-Embed-v2 [30]	Mistral [23]	Text	7.9B	0	2.46M	75.3	70.7	61.7	58.1	56.5	34.2	43.0	38.6
CLIP [47]	Scratch	Image	428M	400M	0	54.6	38.6	38.1	29.7	45.3	20.6	23.2	17.6
DSE [37]	Phi3V [1]	Image	4.2B	0	5.61M	72.7	68.5	73.0	67.2	67.4	49.6	55.5	47.7
VisRAG-Ret [66]	MiniCPM-V [63]	Image	3.4B	0	240K	**87.2***	75.5*	74.3*	68.4*	71.9*	51.7*	56.4	44.5
Trained on OpenDocVQA													
Phi3 [1]	Phi3V [1]	Text	4B	0	41K	72.5	65.3	53.3	48.4	53.2*	33.0*	40.5*	32.0*
VDocRetriever†	Phi3V [1]	Image	4.2B	0	41K	84.2+11.7	74.8+9.5	71.0+17.7	65.1+16.7	66.8*+13.6	52.8*+19.8	48.4*+7.9	41.0*+9.0
VDocRetriever	Phi3V [1]	Image	4.2B	500K	41K	86.0+1.8	**76.4**+1.6	**77.3**+6.3	**73.3**+8.2	**72.9***+6.1	**55.5***+2.7	**57.7***+9.3	**50.9***+9.9

Table 3. Retrieval results under the single- (Single) and all-pool (All) settings. * indicates performance on test data for which corresponding training samples are available. All other results represent zero-shot performance. Init, FT, and PT denote the initialization model, fine-tuning, and pre-training, respectively. Performance gains in green and blue are compared to the base LLM and VDocRetirver†, respectively.

- **Experiments**

Generator	Retriever	Docs	ChartQA Single	ChartQA All	SlideVQA Single	SlideVQA All	InfoVQA Single	InfoVQA All	DUDE Single	DUDE All
Closed-book										
Phi3	–	–	20.0	20.0	20.3	20.3	34.9*	34.9*	23.1*	23.1*
Text-based RAG										
Phi3	Phi3	Text	28.0	28.0	28.6	28.0	40.5*	39.1*	40.1*	35.7*
Phi3	Gold	Text	36.6	36.6	27.8	27.8	45.6*	45.6*	55.9*	55.9*
VDocRAG (Ours)										
VDocGenerator	VDocRetriever	Image	**52.0** +24.0	**48.0** +20.0	**44.2** +15.6	**42.0** +14.0	**56.2*** +15.7	**49.2*** +10.1	**48.5*** +8.4	**44.0*** +8.3
VDocGenerator	Gold	Image	74.0	74.0	56.4	56.4	64.6*	64.6*	66.4*	66.4*

Table 4. DocumentVQA results. All models are fine-tuned on OpenDocVQA. The results marked with * denote performance on unseen test samples, and the other results represent zero-shot performance. The performance gain in green is compared to the text-based RAG that has the same base LLM. Gold knows the ground-truth documents. Models answer the question based on the top three retrieval results.

- **Experiments**

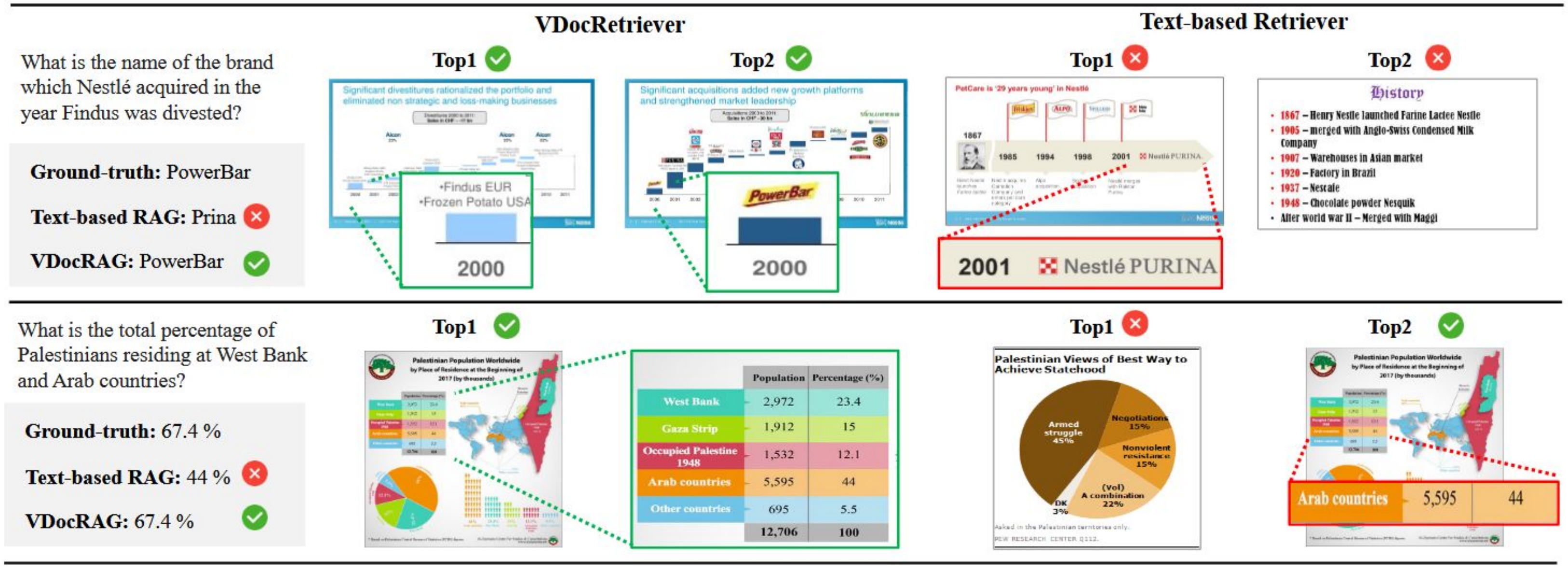


Figure 6. Qualitative results of VDocRAG compared to text-based RAG.