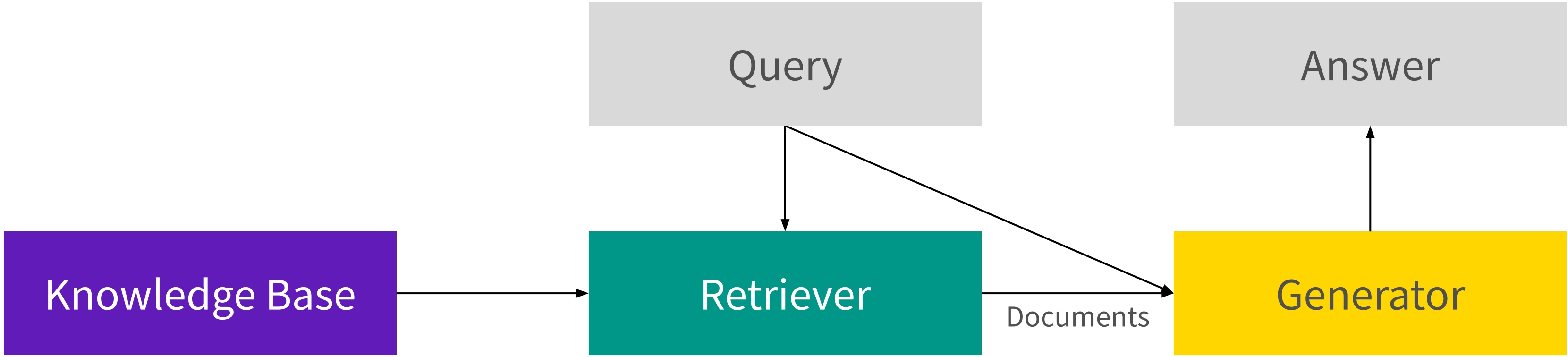
# VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents

Shi Yu[1*] (Speaker), Chaoyue Tang[2*], Bokai Xu[2*], Junbo Cui[2*], Junhao Ran[3], Yukun Yan[1], Zhenghao Liu[4], Shuo Wang[1], Xu Han[1], Zhiyuan Liu[1], Maosong Sun[1]
[1]Tsinghua University [2]ModelBest Inc. [3]Rice University [4]Northeastern University
* equal contribution

*yus21@mails.tsinghua.edu.cn*

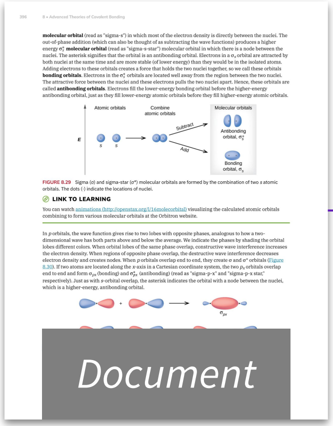
# Background: Retrieval-augmented Generation (RAG)




- RAG supplements the LLM (generator) with external information
- Text snippets are usually the processing units of traditional RAG

# RAG for Real-world Multi-modality Documents

- Real-world documents are often presented in mixed modality, where texts, images, (tables, …) are interleaved on a page with a specific layout
- *Document parsing* is introduced in RAG frameworks to extract texts




*Parsing*

*Text Snippets*

Molecular orbital …

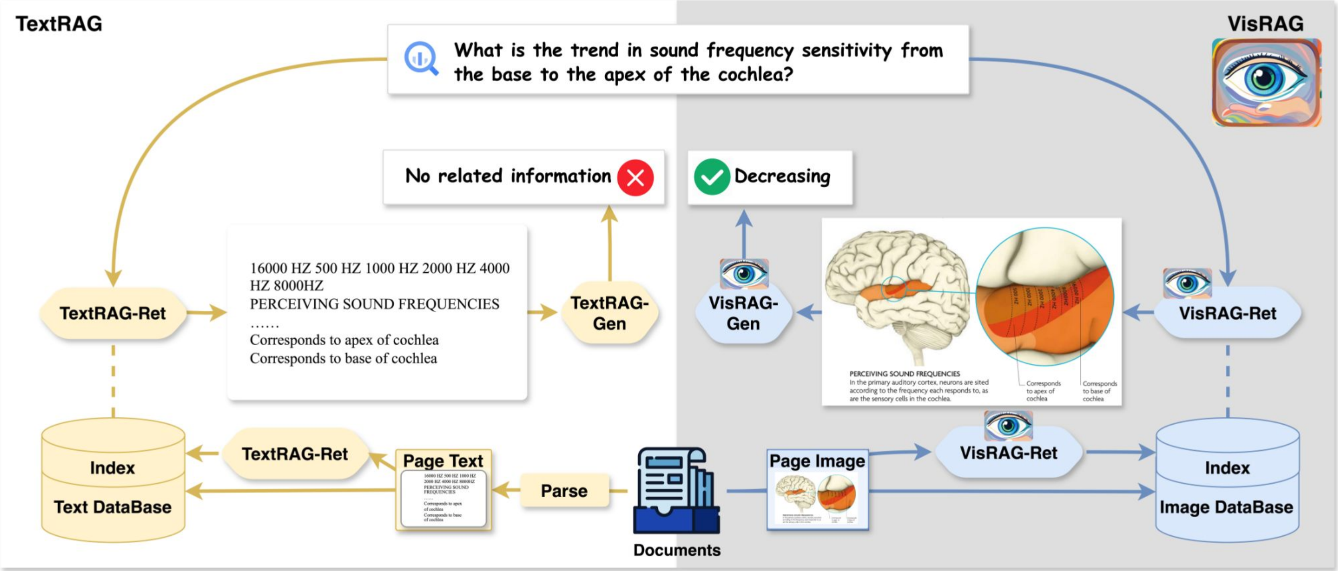
For retrieval or Generation

*harm*

🙁 ***Information Loss**:*
*Visual info (images, layout, …) discarded*
*Potential errors introduced*

# VisRAG: Parsing-free Vision-based RAG

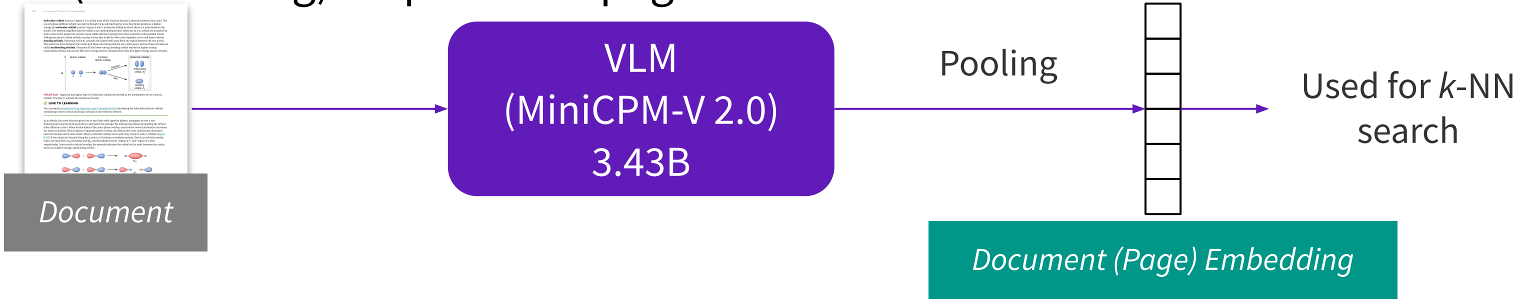
- Parsing-free: Use *document image* as the processing unit
- Vision-based: Use *VLMs* rather than LLMs for retrieval & generation




ICLR 2025

# VisRAG: Method

- Retrieval (Embedding): Represent a page of the document with a VLM




optimized using: $l(q, d^+, D^-) = -\log \frac{\exp(s(q,d^+)/\tau)}{\exp(s(q,d^+)/\tau) + \sum_{d^- \in D^-} \exp(s(q,d^-)/\tau)}$

- Generation:

# VisRAG: Data

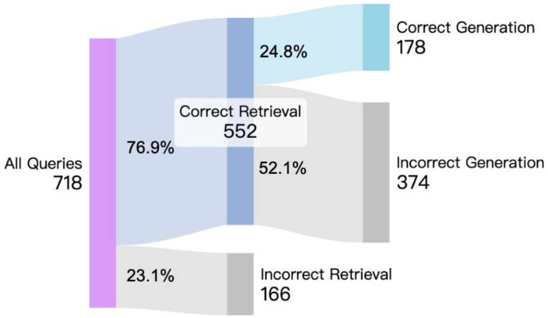
- Synthetic data *for training*
  - Collect PDFs from the Web
  - Prompt GPT-4o on document images to generate queries
- Open-source data from VQA datasets *for training & evaluation*
  - We collect VQA question-document pairs and pool the documents to build retrieval corpus
  - Filtering
    - Some queries from the VQA datasets are *context-dependent*: *Where was the conference held*?
    - We prompt GPT-4o with demonstrations to filter them out
- Data Statistics

| Source | Document Type | Train | Evaluation | | |
|---|---|---|---|---|---|
| | | # Q-D Pairs | # Q (% Preserved) | # D | # Pos. D per Q |
| ArXivQA (2024b) | Arxiv Figures | 25,856 | 816 (8%) | 8,066 | 1.00 |
| ChartQA (2022) | Charts | 4,224 | 63 (5%) | 500 | 1.00 |
| MP-DocVQA (2023) | Industrial Documents | 10,624 | 591 (11%) | 741 | 1.00 |
| InfoVQA (2022) | Infographics | 17,664 | 718 (26%) | 459 | 1.00 |
| PlotQA (2020) | Scientific Plots | 56,192 | 863 (4%) | 9,593 | 1.00 |
| SlideVQA (2023) | Slide Decks | 8,192 | 556 (25%) | 1,284 | 1.26 |
| Synthetic | Various | 239,358 | - | - | - |

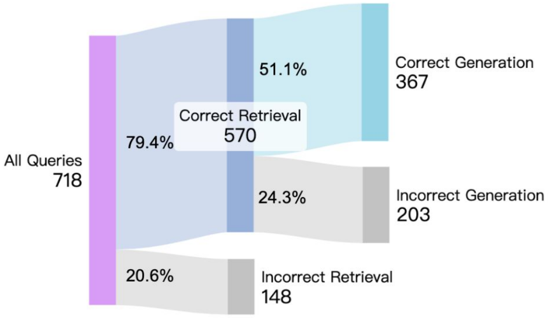
# VisRAG: Overall Results

- The *cascade effect* (retrieval+generation) results in significant performance boost over text-based RAG (TextRAG)

On InfographicsVQA

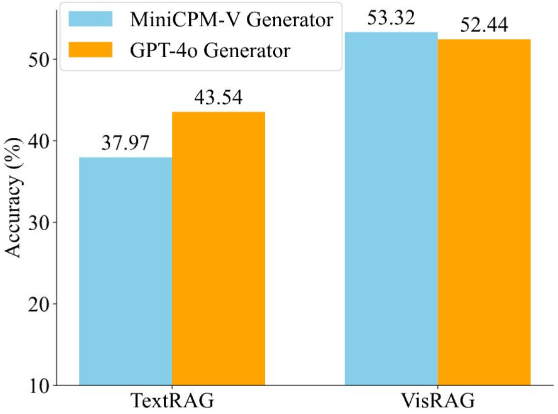



(a) TextRAG with MiniCPM (OCR) as the retriever and MiniCPM-V 2.6 (OCR) as the generator.




(b) VisRAG with VisRAG-Ret as the retriever and MiniCPM-V 2.6 as the generator.

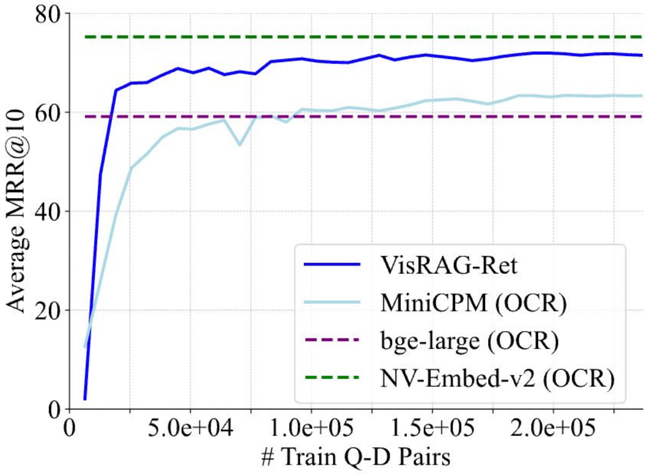
Avg. Performance on all Datasets



# VisRAG: Analysis

- Training data efficiency
  - VisRAG-Ret (the retriever, 3.4B) surpasses bge-large (OCR) after trained on ~20K synthetic data
  - … and achieves 95% of the performance of NV-Embed-v2 (7.9B) after trained on 240K (all) synthetic data
  - Bge-large and NV-Embed-v2 are trained on *millions of* curated query-document pairs
  - *Capturing multi-modal information* is more effective and efficient than merely *increasing training data and model parameters* but relying solely on the text modality

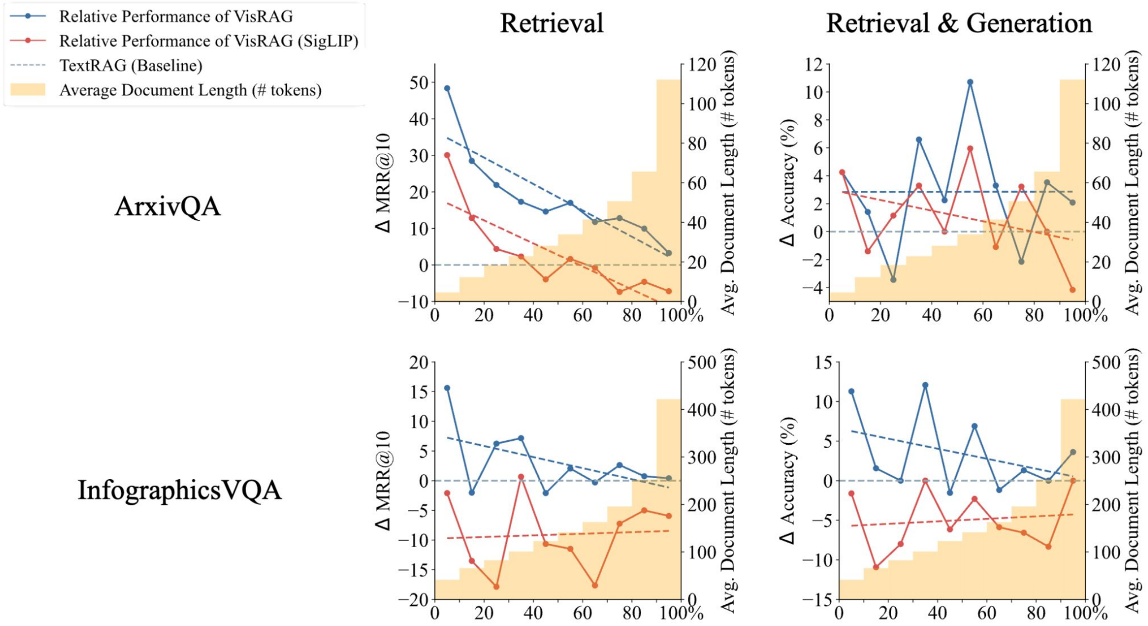
All runs in the out-of-domain setting



# VisRAG: Analysis

- VisRAG performs better than TextRAG on all subsets of documents including text-emphasized ones

Vision-emphasized ← → Text-emphasized



# Conclusion

- VisRAG shows that building a *parsing-free*, *vision-based* RAG pipeline is possible and performs better than text-based RAG (TextRAG) pipelines
- Training the retriever of VisRAG is more *data-efficient* and it *generalizes better* than text retrieval models
- VisRAG is *a more effective and efficient RAG pipeline* than TextRAG for multi-modality documents
- Check out our code and models at https://github.com/openbmb/visrag