A Picture is Worth a Thousand Words? An Empirical Study of Aggregation Strategies for Visual Financial Document Retrieval

Ho Hung Lim and Yi Yang
The Hong Kong University of Science and Technology
limhohung@ust.hk, imyiyang@ust.hk

Abstract

Visual RAG has offered an alternative to traditional RAG. It treats documents as images and uses vision encoders to obtain vision patch tokens. However, hundreds of patch tokens per document create retrieval and storage challenges in a vector database. Practical deployment requires aggregating them into a single vector. This raises a critical question: does single-vector aggregation lose key information in financial documents? We develop a diagnostic benchmark using financial documents where changes in single digits can lead to significant semantic shifts. Our experiments show that single-vector aggregation collapses different documents with almost identical vectors. Metrics show that the patch level detects semantic changes, and confirm that aggregation obscures these details. We identify global texture dominance as the root cause. Our findings are consistent across model scales, retrieval-optimized embeddings, and multiple mitigation strategies, highlighting significant risks for single-vector visual document retrieval in financial applications.

1 Introduction

Retrieval-Augmented Generation (RAG) systems (Lewis et al., 2020) are widely used in the financial domain for analyzing complex financial documents such as annual reports and 10-K reports (Dadopoulos et al., 2025; Kim et al., 2025). They commonly use PDF parsing or Optical Character Recognition (OCR) for extracting document elements and converting them into a linear text sequence (Si et al., 2025; Dadopoulos et al., 2025). However, documents with tables displaying rows and columns are forced to form a linear text sequence that destroys the table structure and breaks row-column alignments. Because document structural context is destroyed during this process, the document retrieval performance decreases (Yu et al., 2025b).

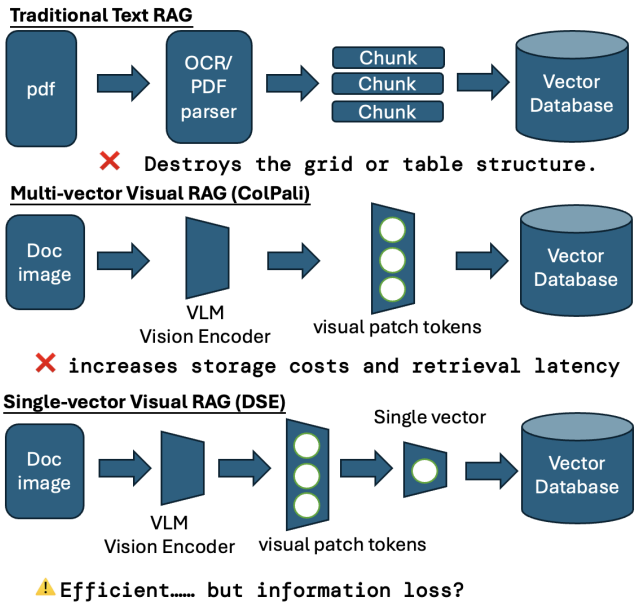

Figure 1: Overview of document retrieval paradigms.

Recent advancements leverage the superior image processing capabilities of Vision-Language Models (VLMs) to extract document content for OCR and retrieval (Kim et al., 2022; Ma et al., 2024; Faysse et al., 2024; Yu et al., 2025a). DeepSeek-OCR (Wei et al., 2025) proves the efficacy of encoding long textual context directly into visual tokens that can achieve high accuracy in OCR tasks. Similarly, ColPali (Faysse et al., 2024) uses VLMs for the representation of document page images as a sequence of visual patch tokens for document retrieval. Despite its success in effectively utilizing document information conveyed by visual patch features, there is consequently higher storage cost due to embedding hundreds of vectors for document page images in the vector database setup; this is a typical characteristic of multi-vector storage in retrieval systems (MacAvaney et al., 2025). For practical deployment, dense retrieval methods such as DSE (Ma et al., 2024) compress these visual tokens into a single vector for efficient storage and faster document retrieval.

Findings of the Association for Computational Linguistics: ACL 2026, pages 15820–15829
July 2-7, 2026

Table-centric financial documents differ fundamentally from generic visual texts in that their most salient information is often encoded in sparse numerical values and key financial entities, rather than in dominant visual structures. This raises an important research question: **When we compress visual patch tokens from a dense financial document into a single vector through aggregation, do we lose key numeric or semantic information?** We hypothesize that this loss is particularly severe for financial documents, where dense numerical data appear against dominant background layouts. Although a single digit changes the semantic meaning, the visual signal of that digit is tiny. As a result, aggregation strategies often prioritize the large background features and smooth over these sparse numeric details. Therefore, we address two critical questions in the financial domain:

1. **Severity**: How severe is this information loss for financial documents, where a single digit change (e.g., \$1.2M $\rightarrow$ \$7.2M) or a textual shift (e.g., changing a date) represents a significant semantic difference?
2. **Mechanism**: What causes this failure?

To answer these questions, we build a diagnostic benchmark. We find that single-vector aggregation fails to preserve the fine-grained details detected by the encoder. We make the following contributions:

- **Aggregation Failure.** We show that **single-vector aggregation** causes different document images to look almost identical (similarity > 0.99), consistently across model scales (7B to 32B) and retrieval-optimized embeddings.
- **Diagnostic Proof.** We use **MinPatch** to find the signal. It recovers the score (Similarity ≈ 0.51). This proves the encoder features are good, but the aggregation ruins them. Simple mitigation strategies also fail to recover the signal.
- **Mechanistic Explanation.** We find the cause is **global texture dominance**. Our analysis shows that single-vector aggregation focuses on the background layout or grid lines instead of the table data.

2 Related Work

2.1 VLMs and Compression

Modern Vision-Language Models (VLMs) like LLaVA (Liu et al., 2023) and Qwen2.5-VL (Bai et al., 2025) encode images into sequences of visual patch tokens through their vision encoder. Recently, DeepSeek-OCR (Wei et al., 2025) achieves remarkable OCR performance by effectively compressing extremely long text contexts into visual token representations. It proves that sequences of visual patch tokens can preserve fine-grained visual details. Although these patch representations are highly effective in generation tasks, they are impractical in the context of retrieval due to the cost of storing hundreds of tokens for every document in a large-scale vector storage system.

2.2 Visual Document Retrieval

Existing methods for visual document retrieval are generally of two types: single-vector and multi-vector approaches. Single-vector approaches, like DSE (Ma et al., 2024), compress visual tokens into a single vector through aggregation (mean pooling or [CLS] token) for efficient retrieval. Though scalable, they fail to retain details of complex layouts. Multi-vector approaches, represented by ColPali (Faysse et al., 2024), use late interaction (MaxSim) over visual patches. ColPali maintains vision tokens for all patches, retaining fine-grained visual details at the cost of storage and latency overhead.

3 Methodology

3.1 Sensitivity Analysis

In order to evaluate whether the vision embeddings accurately capture the semantic meaning within financial document images, we divide our sensitivity analysis into two categories: numeric sensitivity and text sensitivity.

3.2 Numeric Sensitivity

We investigate numeric sensitivity using two distinct perturbation levels.

Condition	Original	Counterfactual
Micro-Semantic	5.21	5.29
(Small Change)	19.65%	19.54%
	10,520	10,526
Macro-Semantic	5.21	9.99
(Large Change)	11.9	88.8
	13,499	99,999

Table 1: Examples of numeric sensitivity strategies.

3.3 Text Sensitivity

We extract the question-answer pairs from the dataset ground truth and manually locate the visual span of the answer within the document image. We then adapt the *semantic occlusion* technique

proposed by Zeiler and Fergus (2014), covering the answer span with the exact background color.

Type	Original	Occluded
Revenue	Revenue increased by **$1.4 billion**.	Revenue increased by [MASK].
Date	Period ending **Dec 31**.	Period ending [MASK].

Table 2: Example of text sensitivity strategies. Ground-truth answers are occluded with the background color (symbolized as [MASK]).

3.4 Structure Analysis

We conduct a **visual attention analysis** to test whether vision embeddings focus on the background layout (including non-informative headers) or the tabular data. For this analysis, we generate three image versions.

- **Reference:** The original document image.
- **Signal Image:** We preserve the table content while filling the entire surrounding area (including headers, footers, and margins) with the document's dominant background color. This isolates the table structure and data from any page context.
- **Noise Image:** We fill the internal table area with the identical background color, effectively "erasing" the table. This leaves only the document template (e.g., logos, pagination) visible against a uniform background.

We measure the cosine similarity of the signal and noise images against the reference to determine which component dominates the single vector.

3.5 Retrieval Mechanisms

To quantify the information loss caused by vector compression, we compare five scoring mechanisms. Let $V_A = \{v_1^A, ..., v_n^A\}$ and $V_B = \{v_1^B, ..., v_n^B\}$ be the patch embedding sequences for the original and counterfactual documents, respectively.

Aggregation: Aggregating all patches into a single vector before computing similarity:

- **Mean Pooling:** $S_{mean} = \cos(\frac{1}{n}\sum v_i^A, \frac{1}{n}\sum v_i^B)$. Averages all patch vectors into a single vector.
- **Max Pooling:** $S_{max} = \cos(\max(V_A), \max(V_B))$. Selects the dimension-wise maximum features.

Late Interaction: Computing similarity at the patch level:

- **MaxSim** (Khattab and Zaharia, 2020): $S_{ms} = \frac{1}{n}\sum_i \max_j \cos(v_i^A, v_j^B)$. Averages the best-match similarity for every patch.
- **MeanPatch:** $S_{mp} = \frac{1}{n}\sum_i \cos(v_i^A, v_i^B)$. Averages the similarity of spatially aligned patches.
- **MinPatch:** $S_{min} = \min_i \cos(v_i^A, v_i^B)$. Identifies the single worst similarity score to isolate local semantic deviations. We use MinPatch as a diagnostic probe, not as a practical retrieval metric.

3.6 Mitigation Strategies

To explore whether simple modifications to the aggregation process can mitigate global texture dominance, we design three alternative aggregation strategies: variance-weighted pooling, attention-guided pooling, and top-k patch removal. Details are provided in Appendix C.

4 Experiments

Experiment	FinQA	TAT-DQA	Total
Sensitivity Analysis			
Micro-Semantic	100	100	200
Macro-Semantic	100	100	200
Text Sensitivity	100	100	200
Visual Attention Analysis			
Signal (Table Only)	100	100	200
Noise (Context Only)	100	100	200

Table 3: Dataset statistics. We construct a balanced diagnostic set (N=200 pairs) for sensitivity analysis and visual attention analysis.

4.1 Dataset Setup

We test on two financial datasets: **FinQA** (Chen et al., 2021) and **TAT-DQA** (Zhu et al., 2021, 2022). Details on image extraction and masking procedure are in Appendix A.

4.2 Models

Following recent work on analyzing VLM information loss (Li et al., 2025), we test the vision encoders of five standard VLMs: **Qwen2.5-VL-7B/32B** (Bai et al., 2025), **LLaVA-v1.5** (Liu et al., 2023), **Phi-3.5-Vision** (Abdin et al., 2024), and **DeepSeek-DeepEncoder** (Wei et al., 2025). We extract the sequence of visual patches output by the projection layer for each model. To evaluate whether retrieval-specific training resolves the aggregation failure, we additionally test two embedding models: **Qwen3-VL-Embedding-8B** (Li et al., 2026) and **GME-Qwen2-VL-7B-Instruct** (Zhang et al., 2024). Full implementation details are provided in Appendix B.

	FinQA					TAT-DQA				
Model	**Mean**	**Max**	**MaxSim**	**MeanP**	**MinP**	**Mean**	**Max**	**MaxSim**	**MeanP**	**MinP**
Micro-Semantic Sensitivity (Values close to 1.0 indicate blindness)										
Qwen2.5-VL-7B	1.0000	0.9996	0.9982	0.9975	**0.7313**	1.0000	1.0000	0.9995	0.9994	**0.6959**
Qwen2.5-VL-32B	0.9999	0.9993	0.9979	0.9969	**0.7123**	1.0000	0.9999	0.9995	0.9994	**0.7107**
LLaVA-v1.5	1.0000	1.0000	0.9992	0.9974	**0.7784**	1.0000	1.0000	0.9996	0.9991	**0.7240**
Phi-3.5-Vision	1.0000	0.9999	0.9980	0.9979	**0.8747**	1.0000	1.0000	0.9997	0.9997	**0.9002**
DeepEncoder	0.9999	0.9997	0.9987	0.9984	**0.9652**	1.0000	0.9999	0.9998	0.9998	**0.9784**
Macro-Semantic Sensitivity										
Qwen2.5-VL-7B	0.9998	0.9985	0.9939	0.9906	**0.5160**	0.9998	0.9998	0.9974	0.9955	**0.5205**
Qwen2.5-VL-32B	0.9999	0.9989	0.9936	0.9899	**0.5289**	1.0000	0.9997	0.9977	0.9960	**0.5635**
LLaVA-v1.5	0.9999	1.0000	0.9955	0.9883	**0.6177**	0.9999	1.0000	0.9981	0.9947	**0.6339**
Phi-3.5-Vision	1.0000	0.9989	0.9925	0.9920	**0.7408**	1.0000	0.9994	0.9974	0.9969	**0.7838**
DeepEncoder	0.9995	0.9985	0.9915	0.9882	**0.8830**	0.9997	0.9997	0.9976	0.9963	**0.9304**
Text Sensitivity										
Qwen2.5-VL-7B	0.9980	0.9914	0.9620	0.9253	**0.1384**	0.9997	0.9987	0.9832	0.9669	**0.1062**
Qwen2.5-VL-32B	0.9982	0.9956	0.9575	0.9183	**0.0875**	0.9997	0.9990	0.9804	0.9655	**0.1118**
LLaVA-v1.5	0.9973	0.9999	0.9778	0.9273	**-0.0903**	0.9994	1.0000	0.9837	0.9530	**0.0105**
Phi-3.5-Vision	0.9993	0.9949	0.9636	0.9461	**0.2401**	0.9997	0.9980	0.9794	0.9711	**0.2789**
DeepEncoder	0.9982	0.9931	0.9637	0.9352	**0.2705**	0.9996	0.9975	0.9819	0.9700	**0.3894**

Table 4: Complete sensitivity benchmark across aggregation strategies.

5 Results

5.1 The Blindness of Single-vector Aggregation in All Sensitivity Tests

As shown in Table 4, both aggregation methods, Mean Pooling and Max Pooling fail completely in both FinQA and TAT-DQA. The similarity scores stay near 1.0 in all our tests. No matter whether we change the text or the numbers, these methods cannot distinguish between the original and the altered document. This confirms that single-vector aggregation strategies are effectively blind to fine-grained semantic changes. In a dense retrieval index, when semantically different documents (e.g., \$1.2M vs. \$7.2M) have similarity scores near 1.0, they occupy the same vector space region. It becomes extremely difficult for queries to reliably rank one above the other, making precise version discrimination highly challenging in financial applications.

5.2 Slight Recovery via Late Interaction

Late-interaction methods such as MaxSim and MeanPatch operate at the patch level, which should theoretically preserve local details. However, as shown in Table 4, these methods provide only marginal improvement: MaxSim and MeanPatch scores remain above 0.99 across all sensitivity tests. This proves that the global texture (background layout, grid lines, and headers) is strong enough to overpower even these local metrics.

5.3 MinPatch Shows the Hidden Signal

Unlike single-vector aggregation, MinPatch successfully distinguishes between small and large errors. In Table 4, similarity scores drop to about 0.71 for micro-semantic changes but fall much further to 0.51 for macro-semantic changes. This proves that the encoder does see the difference. We see similar results in the text sensitivity test. MinPatch scores drop to 0.09 for Qwen2.5-VL-32B, and even become negative for LLaVA. These results confirm that the vision encoder sees the error clearly. It is the aggregation process that obscures the signal. Notably, DeepEncoder consistently shows the highest MinPatch scores (least sensitive), likely because its OCR-optimized encoder learns representations invariant to pixel-level variations, making its aggregated representation less sensitive to single-digit changes compared to general vision encoders.

	FinQA	
Sensitivity Analysis	**Qwen3-VL Embedding-8B**	**GME-Qwen2-VL-7B-Instruct**
Micro-Semantic	0.9992	0.9970
Macro-Semantic	0.9976	0.9906
Text Sensitivity	0.9799	0.9363

Table 5: Evaluation of retrieval-optimized embedding models on sensitivity benchmarks (FinQA).

	TAT-DQA	
Sensitivity Analysis	**Qwen3-VL Embedding-8B**	**GME-Qwen2-VL-7B-Instruct**
Micro-Semantic	0.9999	0.9991
Macro-Semantic	0.9997	0.9906
Text Sensitivity	0.9932	0.9091

Table 6: Evaluation of retrieval-optimized embedding models on sensitivity benchmarks (TAT-DQA).

5.4 Retrieval-Optimized Embedding Models

A natural question is whether embedding models specifically trained for retrieval tasks can overcome the aggregation failure. As shown in Table 5 and Table 6, both Qwen3-VL-Embedding-8B and GME-Qwen2-VL-7B-Instruct exhibit the same blindness (similarity near 1.0). This confirms that the failure is inherent to single-vector representations, regardless of training objectives.

5.5 Preliminary Mitigation Attempts

To explore whether simple aggregation strategies can mitigate global texture dominance, we tested three approaches: variance-weighted pooling (VarWgt), attention-guided pooling (AttnGd), and top-k patch removal (TopK-R). As shown in Table 7, all three methods provide only marginal improvement, with similarity scores remaining above 0.99 across all sensitivity tests. This confirms the severity of global texture dominance: background features are so pervasive that simple re-weighting or filtering strategies cannot recover the fine-grained signals. Combined with our MinPatch results (Table 4), which show similarity as low as 0.51, this demonstrates that the problem is fundamental to single-vector aggregation.

5.6 Layout vs Tabular Data

Since we want to know why single-vector aggregation fails on the sensitivity test, we perform a visual attention analysis to see whether the vision embeddings focus on the background layout (including non-informative headers) or the tabular data. The performance gap between the similarity scores of the background layout and the tabular data helps us identify what the single vector pays attention to. As shown in Table 8, both Qwen2.5-VL-7B and Qwen2.5-VL-32B show a large bias ($\Delta = +0.22$ and $+0.24$, respectively) in FinQA. This indicates the single vector focuses on the background layout rather than the tabular data. On the other hand, DeepSeek-DeepEncoder shows a lower bias ($\Delta = +0.08$) in FinQA. This suggests the single vector of DeepSeek focuses less on the background layout but still not enough to identify the changes. In TAT-DQA, the gap becomes negligible ($\Delta < 0.05$) due to global texture dominance. This prevents the single vector from distinguishing between table content and background layout, effectively blinding the single-vector representation to the actual data within visually complex tables.

6 Conclusion

In this paper, we develop a diagnostic benchmark to analyze the reliability of single-vector aggregation in visual document retrieval for table-centric financial documents. We find that although VLMs' vision encoders successfully capture fine-grained numeric and textual details, aggregation methods such as mean pooling obscure these signals due to global texture dominance. Our findings are robust across varying model scales, specialized retrieval-optimized architectures, and multiple mitigation strategies, all of which fail to recover the lost signals. These results highlight significant risks for single-vector visual document retrieval in financial applications and suggest that multi-vector retrieval approaches or learned aggregation methods are necessary for practical deployment.

Limitations

Our diagnostic benchmark focuses on table-centric financial documents from two datasets (FinQA and TAT-DQA). While these represent common financial document formats, other types such as invoices, balance sheets with different layouts, or handwritten financial notes may exhibit different levels of global texture dominance. Extending the benchmark to cover a broader range of financial document types would strengthen the generalizability of our findings. Additionally, the findings on global texture dominance may not transfer to other domains such as natural images, which have fundamentally different visual characteristics (see Appendix E). Furthermore, our benchmark may not cover all real-world semantic changes since we only work on specific numeric and textual perturbations. Future work should explore more diverse perturbation types and larger-scale evaluation across financial document categories. While our study focuses on diagnostic similarity analysis between document pairs, a full retrieval evaluation with ranking metrics (e.g., Recall@k, nDCG) over a large corpus would further contextualize the practical impact and is left for future work.

Acknowledgments

This work was conducted as part of a HKUST Frontier Technology Research for Joint Institutes with Industry (FTRIS) project and was supported by WeBank under Grant No. WEB25BM01.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michail Dadopoulos, Anestis Ladas, Stratos Moschidis, and Ioannis Negkakis. 2025. Metadata-driven retrieval-augmented generation for financial question answering. *Preprint*, arXiv:2510.24402.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Sejong Kim, Hyunseo Song, Hyunwoo Seo, and Hyunjun Kim. 2025. Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems. *Preprint*, arXiv:2503.15191.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Chen Keqin, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulić, and Anders Søgaard. 2025. Lost in embeddings: Information loss in vision–language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22676–22693, Suzhou, China. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Sean MacAvaney, Antonio Mallia, and Nicola Tonellotto. 2025. Efficient constant-space multi-vector retrieval. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III*, page 237–245, Berlin, Heidelberg. Springer-Verlag.

Jacob Si, Mike Qu, Michelle Lee, and Yingzhen Li. 2025. Tabrag: Tabular document retrieval via structured language representations. *Preprint*, arXiv:2511.06582.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025a. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*.

Xiaohan Yu, Pu Jian, and Chong Chen. 2025b. TableRAG: A retrieval augmented generation framework for heterogeneous document reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14063–14082, Suzhou, China. Association for Computational Linguistics.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4857–4866. ACM.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Dataset Details

FinQA. FinQA (Chen et al., 2021) focuses on numerical reasoning in financial reports. We create the image set by manually taking screenshots of the test documents[1] to simulate real-world viewing conditions.

TAT-DQA. TAT-DQA (Zhu et al., 2022), an extension of the TAT-QA dataset (Zhu et al., 2021), is considered harder than FinQA. It contains multi-page documents with dense financial tables. We extract these images directly from the source PDFs.

Masking Procedure. Unlike automated bounding boxes which often suffer from localization errors, we manually verify the visual boundaries of the table region for each document. This ensures pixel-perfect alignment with the table's semantic content, avoiding partial occlusions or leftover artifacts common in automated methods.

B Implementation Details

Setup. We implemented all models using the HuggingFace Transformers library (Wolf et al., 2020) on a single NVIDIA RTX 5880 Ada Generation GPU (48GB). To ensure a fair comparison, we utilized the official pre-trained weights for all architectures (e.g., `Qwen/Qwen2.5-VL-7B-Instruct`). For DeepEncoder, we adapted the implementation from `https://github.com/Volkopat/VLM-Optical-Encoder`. Embedding similarity was computed using **Cosine Similarity**, and all input images were resized to the model's default resolution to prevent resizing artifacts.

[1] `https://finqasite.github.io/explore.html`

C Mitigation Strategy Details

To explore whether simple modifications to the aggregation process can preserve fine-grained signals, we design three straightforward baseline strategies. We emphasize that these strategies are not proposed as optimal solutions, but rather as simple probes to test whether straightforward aggregation modifications can recover fine-grained signals lost during single-vector compression.

- **Variance-Weighted Pooling (VarWgt):** Assigns higher weights to patches with greater variance across embedding dimensions, under the hypothesis that informative patches exhibit higher variance than repetitive background patches.
- **Attention-Guided Pooling (AttnGd):** Computes a patch-level self-similarity matrix via normalized dot products among patch embeddings, and uses the standard deviation of each patch's similarity scores as a proxy for patch informativeness. Patches with low diversity are down-weighted under the hypothesis that uniformly similar patches correspond to repetitive background regions.
- **Top-k Removal (TopK-R):** Computes cosine similarity between spatially aligned patch pairs from the original and perturbed documents, removes the top-k most similar aligned patches, and mean-pools the remaining patches. We use a fixed $k = 50$ as a simple heuristic. Note that this corresponds to different removal proportions across encoders (around 9–19% depending on model).

	FinQA			**TAT-DQA**		
Model	**VarWgt**	**AttnGd**	**TopK-R**	**VarWgt**	**AttnGd**	**TopK-R**
Micro-Semantic Sensitivity (Values close to 1.0 indicate blindness)						
Qwen2.5-VL-7B	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
Qwen2.5-VL-32B	0.9997	0.9999	0.9999	1.0000	1.0000	1.0000
LLaVA-v1.5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Phi-3.5-Vision	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
DeepEncoder	0.9999	0.9999	0.9999	1.0000	1.0000	1.0000
Macro-Semantic Sensitivity						
Qwen2.5-VL-7B	0.9997	0.9998	0.9998	0.9996	0.9999	0.9999
Qwen2.5-VL-32B	0.9997	0.9998	0.9998	0.9999	1.0000	1.0000
LLaVA-v1.5	1.0000	0.9999	0.9999	0.9999	0.9999	0.9999
Phi-3.5-Vision	0.9999	0.9999	0.9999	0.9999	1.0000	1.0000
DeepEncoder	0.9994	0.9995	0.9994	0.9998	0.9997	0.9996
Text Sensitivity						
Qwen2.5-VL-7B	0.9962	0.9982	0.9979	0.9992	0.9997	0.9832
Qwen2.5-VL-32B	0.9972	0.9980	0.9980	0.9996	0.9996	0.9996
LLaVA-v1.5	0.9998	0.9974	0.9981	0.9999	0.9992	0.9992
Phi-3.5-Vision	0.9989	0.9990	0.9988	0.9995	0.9996	0.9995
DeepEncoder	0.9980	0.9982	0.9978	0.9996	0.9996	0.9995

Table 7: Evaluation of mitigation strategies on sensitivity benchmarks.

D Visual Attention Analysis

Model	FinQA			TAT-DQA		
	Sim to Data (Table Only)	**Sim to Layout**	**Gap**	**Sim to Data** (Table Only)	**Sim to Layout**	**Gap**
STRATEGY: MEAN POOLING (Values close to 1.0 indicate blindness)						
Qwen2.5-VL-7B	0.7548	0.9791	**0.2243**	0.9268	0.9464	**0.0196**
Qwen2.5-VL-32B	0.7369	0.9724	**0.2355**	0.9343	0.9222	**-0.0121**
LLaVA-v1.5	0.8040	0.9847	**0.1807**	0.9364	0.9684	**0.0320**
Phi-3.5-Vision	0.7171	0.9911	**0.2740**	0.9190	0.9693	**0.0503**
DeepEncoder	0.9106	0.9866	**0.0760**	0.9222	0.9667	**0.0446**
STRATEGY: MAX POOLING						
Qwen2.5-VL-7B	0.9019	0.9677	0.0658	0.9710	0.9766	0.0056
Qwen2.5-VL-32B	0.9641	0.9822	0.0181	0.9859	0.9809	-0.0051
LLaVA-v1.5	0.9986	0.9997	0.0012	0.9995	0.9997	0.0002
Phi-3.5-Vision	0.9208	0.9854	0.0646	0.9629	0.9781	0.0152
DeepEncoder	0.9626	0.9817	0.0190	0.9762	0.9811	0.0050
STRATEGY: LATE INTERACT (MAXSIM)						
Qwen2.5-VL-7B	0.6062	0.8948	0.2886	0.7948	0.8286	0.0338
Qwen2.5-VL-32B	0.5401	0.8821	0.3420	0.7570	0.7954	0.0384
LLaVA-v1.5	0.6935	0.9418	0.2482	0.8095	0.8810	0.0715
Phi-3.5-Vision	0.6109	0.9129	0.3020	0.7431	0.8297	0.0866
DeepEncoder	0.7584	0.8944	0.1360	0.7995	0.8584	0.0588
STRATEGY: LATE INTERACT (MEANPATCH)						
Qwen2.5-VL-7B	0.3841	0.8231	**0.4390**	0.6171	0.6974	**0.0803**
Qwen2.5-VL-32B	0.3364	0.8007	**0.4643**	0.6081	0.6672	**0.0591**
LLaVA-v1.5	0.3788	0.8346	**0.4558**	0.5177	0.6764	**0.1587**
Phi-3.5-Vision	0.3028	0.8706	**0.5678**	0.5480	0.7150	**0.1670**
DeepEncoder	0.6232	0.8309	**0.2077**	0.6367	0.7398	**0.1031**
STRATEGY: LATE INTERACT (MINPATCH)						
Qwen2.5-VL-7B	-0.1575	0.0083	0.1658	-0.0455	0.0079	0.0534
Qwen2.5-VL-32B	-0.0080	0.0407	0.0487	0.0183	0.0272	0.0089
LLaVA-v1.5	-0.1257	-0.0592	0.0666	-0.1335	-0.1360	-0.0025
Phi-3.5-Vision	-0.1617	0.0494	0.2111	-0.1091	-0.0526	0.0566
DeepEncoder	0.1670	0.1727	0.0058	0.1798	0.1844	0.0046

Table 8: Complete visual attention analysis results across aggregation strategies.

E Sanity Check — Natural Images vs. Financial Documents

Model	Natural Image (Cat vs Dog)	Financial Doc ($1.2 vs $1.3)
Qwen3-VL-Embedding-8B	0.1192	0.9992
GME-Qwen2-VL-7B-Instruct	0.2165	0.9970
Qwen2.5-VL-7B	0.4038	0.9999
Qwen2.5-VL-32B	0.5117	0.9999

Table 9: Domain comparison: natural images vs. financial documents.

To confirm that the aggregation failure is domain-specific, we compared natural images (cat vs. dog) with financial documents ($1.2M vs. $1.3M) using the same models. Results show natural images maintain discriminability (similarity 0.12–0.51), while financial documents collapse (similarity near 1.0), confirming a domain gap of 0.5–0.9. This validates our hypothesis that sparse numeric signals in background-dominated documents are uniquely vulnerable to single-vector aggregation.