Unified Multimodal Retrieval Framework for Multimodal RAG

Haitao Huang[1], Tianyi Feng[2], Ruiyan Wang[1], Wei Xiong[2], Fei Huang[2], Zhengxue Cheng[1], Rong Xie[1], and Li Song[1(✉)]

[1] Shanghai Jiao Tong University, Shanghai, China
{huanghaitao,rwang0627,zxcheng,xierong,song_li}@sjtu.edu.cn
[2] LongShine Technology Group Co., Ltd., China
{fengtianyi,xiongwei,huangfei}@longshine.com

Abstract. Retrieval-Augmented Generation (RAG) mitigates hallucination in large language models but remains limited for text–image mixed documents due to modality separation, OCR-induced semantic fragmentation, and cross-modal similarity inconsistency. We present a unified multimodal retrieval framework that features three key innovations: a unified multimodal encoder to eliminate modality barriers, a post-encoding residual fusion mechanism to preserve unimodal consistency while capturing cross-modal interactions, and a scaled training strategy to correct modality learning imbalance. This framework maps text, images, and text–image units into a shared semantic space, enabling direct retrieval over coherent multimodal chunks with text queries. Experiments on six benchmarks (ArxivQA, ChartQA, DocVQA, InfoVQA, PlotQA, SlideVQA) show state-of-the-art results: our 3.4B model achieves an average +3.99/+4.27 absolute points in Recall@1/Recall@3 over the strongest baseline (GME) and outperforms larger alternatives (e.g., 8.4B E5-V), validating that unified encoding and scaled training effectively alleviate modality separation, similarity inconsistency, and system complexity for multimodal RAG.

Keywords: Unified Multimodal Retrieval · Retrieval-Augmented Generation (RAG) · Scaled Training Strategy

1 Introduction

Large Language Models (LLMs) have exceptional capabilities but suffer from hallucination and parametric knowledge limitations. Retrieval-Augmented Generation (RAG) mitigates this via external knowledge integration, yet traditional pipelines focus solely on textual content [17]. Real-world knowledge often exists as text-image mixed documents with coherent semantic links, while current OCR-based extraction introduces errors, semantic loss, and overlooks critical visual information.

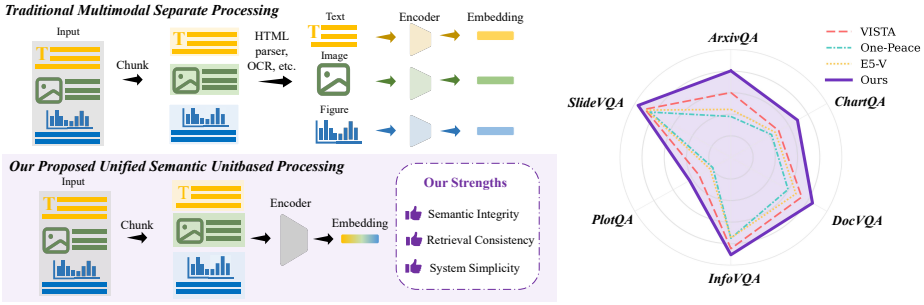

Fig. 1. Unified multimodal retrieval overview. Left: traditional separate text/image pipelines fragment semantics and induce cross-modal inconsistency. Our unified encoder treats text–image chunks as coherent units and supports direct text-to-multimodal retrieval. Right: a radar summary of Recall@1 across six benchmarks (higher is better). Detailed results are reported in Table 2.

Multimedia applications have advanced multimodal retrieval, where unified retrievers outperform separated pipelines. However, existing work focuses on natural images and lacks support for text-image mixed documents.

Current RAG implementations face three critical challenges: (1) semantic fragmentation from separating text-image documents; (2) retrieval inconsistency due to the absence of unified similarity standards; and (3) increased system complexity from multiple independent retrieval pathways.

To address these limitations, we propose a unified multimodal retrieval framework that processes text-image mixed data as coherent semantic units, enabling direct retrieval from multimodal chunks via text queries (as shown in Fig. 1). Our approach comprises a multimodal unified encoding architecture, an effective semantic fusion mechanism, and a balanced training strategy. Experimental validation demonstrates significant performance improvements across multiple benchmarks while simplifying system architecture.

Our key contributions can be summarized as follows:

- We propose a unified framework for text-query-driven tri-modal retrieval that processes pure text, pure image, and text–image mixed chunks within a single semantic space, addressing modality separation issues in traditional RAG systems.
- We design a post-encoding residual fusion module that preserves unimodal consistency while learning cross-modal interactions, and propose a scaled training strategy that compensates for capacity disparity and optimization bias.
- We demonstrate that training innovation can overcome base model limitations, achieving superior performance with smaller models and providing practical guidance for efficient multimodal retrieval system design.

2 Related Work

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) mitigates hallucination in large language models by integrating external knowledge. Dense Passage Retrieval (DPR) [7] pioneered dense retrieval with dual encoders, while Fusion-in-Decoder (FiD) [4] enhances generation through passage fusion.

However, existing RAG methods primarily target textual content. For multimodal documents, OCR-based text extraction introduces errors and disrupts text-image semantic relationships. VisRAG [19] employs VLMs to embed documents as images, achieving $25-39\%$ improvements over text-based RAG, but lacks unified capabilities across pure text, images, and text-image combinations. This paper extends VisRAG with a unified tri-modal encoding architecture and balanced training strategies.

2.2 Multimodal Retrieval

While RAG has advanced text-based retrieval, multimodal representation learning focuses on unifying text and visual information. CLIP [12] and ALIGN [5] advance vision-language representation learning, but focus on natural images rather than document understanding. In document understanding, LayoutLM [16] integrates text, visual, and layout information, and ColPali [2] directly encodes document pages.

Existing methods use modality-specific designs: CLIP optimizes natural image-text matching, ColPali focuses on document image retrieval, and BGE-M3 [1] handles multilingual text. They fail to process mixed modal content under text queries—specifically, uniform modeling of text, image, and text-image chunks, and effective training for cross-modal balance. Critically, uniform sampling causes text-biased optimization with under-learned image and bimodal content, motivating our unified encoder and scaled training strategy.

3 Method

In practical document processing scenarios, RAG systems partition long documents into chunks for retrieval. These chunks are categorized into three types: pure text chunks c^{text}, pure image chunks c^{image}, and text-image mixed chunks c^{bimodal}. Given a text query q, our objective is to retrieve the most relevant content from the chunk collection $\mathcal{C} = \{c^{\text{text}}, c^{\text{image}}, c^{\text{bimodal}}\}$.

Traditional multimodal retrieval methods adopt separated architectures, constructing independent retrievers for text and images. For text-to-mixed retrieval scenarios, separate processing followed by fusion is commonly employed. This specialized design encounters three fundamental challenges: (1) *Modality barrier*: similarity scores from different retrievers lack unified comparison standards; (2)

Semantic fragmentation: separated processing disrupts inherent semantic coupling in mixed-modal chunks; (3) *Training imbalance*: existing methods lack training strategies tailored for tri-modal scenarios.

To address these limitations, our framework maps chunks of all three modalities into the same d-dimensional semantic space through a unified encoding architecture $f_{\boldsymbol{\theta}} : x \to \mathbb{R}^d$, where x encompasses text, image, and mixed inputs (Fig. 2). The retrieval process is uniformly defined as:

$$R(q, c_i) = f_{\boldsymbol{\theta}}(\boldsymbol{q})^T f_{\boldsymbol{\theta}}(\boldsymbol{c}_i). \tag{1}$$

where $\boldsymbol{\theta}$ denotes the set of all trainable parameters of the unified encoding architecture, $f_{\boldsymbol{\theta}}(\boldsymbol{q}) \in \mathbb{R}^d$ and $f_{\boldsymbol{\theta}}(\boldsymbol{c}_i) \in \mathbb{R}^d$ represent the d-dimensional embeddings of query $\boldsymbol{q}$ and candidate chunk $\boldsymbol{c}_i \in \mathcal{C}$, respectively. The inner product quantifies semantic similarity in the unified space, enabling direct comparison and ranking across all modalities.

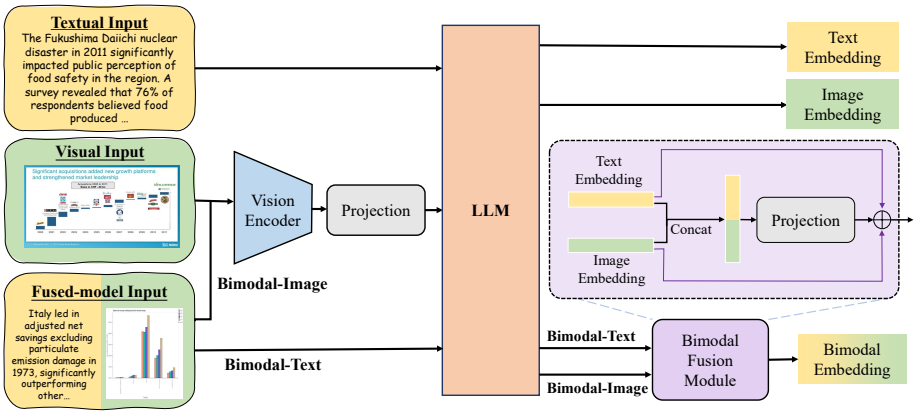

Fig. 2. Overall architecture of the proposed unified multimodal retrieval framework.

3.1 Unified Multimodal Encoder Architecture

We propose a unified multimodal architecture employing a shared encoder that processes all input modalities through a single pre-trained vision-language model. Unlike traditional approaches that employ independent query and document encoders, our framework utilizes a shared vision-language model $\mathcal{E}$ to process both queries and documents simultaneously, ensuring identical semantic spaces. Given an input x (text, image, or multimodal), the encoder performs encoding based on the modality indicator r:

$$\mathbf{H} = \mathcal{E}(x, r), \tag{2}$$

where $r \in \{r_{\text{text}}, r_{\text{img}}\}$ denotes the input modality, and $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_s]$ represents the hidden state sequence.

Let d denote the content within a document chunk; for bimodal chunks, d_{text} and d_{img} denote its textual and visual components, respectively. The encoder adaptively handles three scenarios according to modal composition:

$$\mathbf{v}_d = \begin{cases} \text{Encode}(d_{\text{text}}, r_{\text{text}}) & \text{text} \\ \text{Encode}(d_{\text{img}}, r_{\text{img}}) & \text{image} \\ \text{Fusion}(\text{Encode}(d_{\text{text}}, r_{\text{text}}), \text{Encode}(d_{\text{img}}, r_{\text{img}})) & \text{bimodal} \end{cases} \tag{3}$$

Given the causal attention mechanism, we employ position-weighted average pooling:

$$\mathbf{v} = \sum_{i=1}^{S} w_i \mathbf{h}_i, \tag{4}$$

where S denotes the total length of the hidden state sequence output by the encoder, and $w_i = \frac{i}{\sum_{j=1}^{S} j}$ denotes the position weight for the i-th hidden state vector $\mathbf{h}_i$. All embedding vectors undergo L2 normalization:

$$\mathbf{v}_{\text{norm}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}. \tag{5}$$

The similarity between query q and document d is computed via cosine similarity:

$$s(q, d) = \mathbf{v}_q^T \mathbf{v}_d. \tag{6}$$

The model is optimized using InfoNCE loss:

$$\mathcal{L}(q, d^+, D^-) = -\log \frac{\exp\big(s(q, d^+)/\tau\big)}{\exp\big(s(q, d^+)/\tau\big) + \sum_{d^- \in D^-} \exp\big(s(q, d^-)/\tau\big)}, \tag{7}$$

where q is the text query, d^+ denotes a positive sample, D^- is the set of negative samples, τ is the temperature parameter, and in-batch negative sampling is adopted for training.

3.2 Bimodal Semantic Fusion Mechanism

Semantic representation of bimodal documents is critical for effective multimodal retrieval. Existing early fusion approaches mix modalities before encoding, losing unimodal discriminability, while late fusion methods (e.g., averaging) fail to capture cross-modal interactions. We propose a post-separate-encoding residual fusion mechanism that first encodes each modality independently to preserve unimodal semantics, then learns cross-modal interactions through a learnable fusion layer with residual connections, ensuring both semantic consistency and effective cross-modal alignment.

Our encoding framework adheres to three core design principles: *semantic consistency*, ensuring identical content generates consistent representations; *composability*, enabling complex multimodal representations to be decomposed into unimodal components; and *robustness*, maintaining system functionality when modalities are absent. Following these principles, we employ a separate encoding strategy.

For a bimodal document $(d_{\text{text}}, d_{\text{img}})$, each modality is encoded independently:

$$\mathbf{v}_{\text{text}}^{(bi)} = \text{Normalize}(\text{Pool}(\mathcal{E}_{\text{shared}}(d_{\text{text}}, \emptyset))), \tag{8}$$

$$\mathbf{v}_{\text{img}}^{(bi)} = \text{Normalize}(\text{Pool}(\mathcal{E}_{\text{shared}}(\emptyset, d_{\text{img}}))), \tag{9}$$

where $\mathbf{v}_{\text{text}}^{(bi)}$ and $\mathbf{v}_{\text{img}}^{(bi)}$ are the normalized embeddings of textual and visual components in a bimodal document; $\mathcal{E}_{\text{shared}}$ is the shared unified multimodal encoder; d_{text} and d_{img} are the textual and visual content of a bimodal document; $\emptyset$ is the empty input of the corresponding modality.

The unimodal representations are concatenated and processed through a linear transformation with residual connections to preserve original modal information:

$$\mathbf{h}_{\text{concat}} = [\mathbf{v}_{\text{text}}^{(bi)}; \mathbf{v}_{\text{img}}^{(bi)}] \in \mathbb{R}^{2d}, \tag{10}$$

$$\mathbf{h}_{\text{proj}} = \mathbf{W}_f \mathbf{h}_{\text{concat}} + \mathbf{b}_f \in \mathbb{R}^d, \tag{11}$$

where $\mathbf{W}_f \in \mathbb{R}^{d \times 2d}$ and $\mathbf{b}_f \in \mathbb{R}^d$ are the fusion weight matrix and bias vector, respectively. To prevent information loss, we incorporate a residual connection:

$$\mathbf{h}_{\text{fused}} = \mathbf{h}_{\text{proj}} + \frac{\mathbf{v}_{\text{text}}^{(bi)} + \mathbf{v}_{\text{img}}^{(bi)}}{2}. \tag{12}$$

Finally, L2 normalization ensures the fused representation resides in the same semantic space:

$$\mathbf{v}_{\text{bi}} = \text{Normalize}(\mathbf{h}_{\text{fused}}). \tag{13}$$

This residual linear fusion maintains computational efficiency, effectively captures cross-modal interactions, and yields a single normalized embedding directly comparable to unimodal vectors.

3.3 Scaled Multimodal Training Strategy

In multimodal model training, uniform modal distribution (1:1:1 for text:image:bimodal) consistently yields rapid convergence for text and bimodal tasks, while exhibiting persistently suboptimal performance for image modality. We theoretically analyze this imbalance through three complementary perspectives: (1) *Architectural capacity disparity*: Vision-language models allocate substantially more parameters to textual components than visual ones. For instance, MiniCPM-V 2.0[18] demonstrates a 7:1 parameter ratio (approximately 2.7B language versus 400M visual encoder parameters), creating an inherent

learning capacity asymmetry. (2) *Gradient optimization bias*: Under uniform sampling, the text modality's faster convergence dominates gradient updates, causing the optimizer to prioritize text loss reduction at the expense of visual learning. (3) *Data representation complexity*: Visual information requires hierarchical feature extraction through multiple transformer layers, while text provides direct semantic signals, necessitating more training samples for visual modality to achieve comparable representation quality.

To mitigate these imbalances, we propose a scaled training strategy that compensates for the architectural capacity disparity by increasing image modality exposure. The 2:8:6 ratio (text:image:bimodal) is derived from the parameter ratio analysis: given the 7:1 text-to-visual parameter ratio, we approximately invert this to prioritize visual learning, while maintaining bimodal samples to preserve cross-modal alignment. Specifically, we construct training samples through selective modality removal: *Pure text samples* (12.5%): query-text pairs with empty images; *Pure image samples* (50%): query-image pairs with empty text; *Bimodal samples* (37.5%): complete triplets. Context-dependent queries lacking document specificity are filtered. This ratio balances three objectives: (1) compensating for visual parameter scarcity through increased exposure, (2) maintaining text-visual alignment via bimodal samples, and (3) preserving text quality with minimal text-only samples. Experimental results demonstrate substantial performance improvements and balanced accuracy across modalities, validating both our imbalance analysis and strategy effectiveness.

Our proposed training strategy establishes a fundamental principle: disparities in modal learning capacity require corresponding adjustments in data distribution, providing valuable insights for future multimodal research.

4 Experiments

4.1 Dataset Construction

Our evaluation employs visual question answering (VQA) benchmarks: MP-DocVQA [14] (industrial documents), ArXivQA [8] (academic papers), ChartQA [9](charts), InfographicsVQA [10], PlotQA [11], and SlideVQA [13] (presentation slides) (shown in Table 1). We follow the original datasets' train-test splits, except for MP-DocVQA and InfographicsVQA, where the validation split serves as our evaluation set.

To enable tri-modal retrieval, we augment image-text pairs with textual descriptions generated using Qwen-VL-Max. This augmentation is necessary because many VQA datasets contain image-only candidates without accompanying text, which would prevent fair comparison across text, image, and bimodal chunks in our unified framework. A standardized prompt guides description generation, followed by post-processing including redundancy removal and format standardization.

Table 1. Dataset statistics for the used VQA datasets.

Source	Document Types	Train	Evaluation
ArXivQA	Arxiv Figures	25,856	8,636
ChartQA	Charts	4,224	717
MP-DocVQA	Industrial Documents	10,624	1,878
InfoVQA	Infographics	17,664	2,034
PlotQA	Scientific Plots	56,192	11,306
SlideVQA	Slide Decks	8,192	2,148
Synthetic	Various	122,752	26,719

4.2 Experimental Settings

Training Configuration. We employ MiniCPM-V 2.0 with SigLIP visual encoder and MiniCPM language model. Training uses in-batch negatives with 2 epochs, batch size 32, and temperature 0.02. Position-weighted pooling extracts document representations, optimized with InfoNCE loss on 2 NVIDIA A800 GPUs.

Baselines. We compare against: (1) UMR models VISTA[21] and E5-V[6]; (2) Multimodal models One-Peace[15]; (3) Visual document retrieval VisRAG[19]; (4) Recent multimodal retrievers GME[20] and Jina[3].

Metrics. We use Recall@1 and Recall@3, which are standard metrics in retrieval tasks [7, 6, 20]. Recall@K measures the proportion of queries where at least one relevant document appears in the top-K retrieved results, directly reflecting retrieval quality for RAG applications where users typically examine top-ranked results. We rank queries against unified candidate pools via cosine similarity on ℓ_2-normalized embeddings.

4.3 Results

Table 2 presents comparisons across six datasets, demonstrating our framework's superior performance. Our MiniCPM-V 2.0-based model achieves state-of-the-art performance across all datasets, consistently outperforming existing baselines. Notably, we outperform stronger competitors GME and Jina, which are built on the more powerful Qwen2.5VL foundation model, demonstrating that our unified architecture, learnable fusion mechanism, and theoretically-grounded training strategy effectively compensate for base model scale limitations.

Several key observations emerge from these results. First, our method shows strong performance on visual-intensive tasks: on ChartQA we gain +7.81 Recall@1 over both GME and Jina, indicating better handling of structured charts. PlotQA, which is known for challenging scientific charts, sees +7.48/+6.13 improvements over GME/Jina, highlighting robust visual reasoning. Second, gains extend beyond visual-heavy datasets. On DocVQA (industrial documents) we surpass GME by +1.49 Recall@1, and InfoVQA/SlideVQA gain +1.42/+1.17, demonstrating broad effectiveness across infographics and slides. Third, despite

Table 2. Overall Retrieval Performance in Recall@1/Recall@3. The best retrieval performance in each group is marked in bold, and the second best performance is underlined.

Model	Size	ArxivQA	ChartQA	DocVQA	InfoVQA	PlotQA	SlideVQA
VISTA	0.2B	60.09/64.86	49.65/55.65	73.59/80.99	84.61/90.12	33.03/45.00	89.29/92.40
GME	2.2B	80.02/86.97	57.04/66.11	83.33/91.00	88.89/94.00	35.43/47.42	95.25/97.35
Jina	3.8B	77.86/85.11	61.37/68.90	81.47/89.62	88.69/92.97	36.78/48.54	96.23/97.91
One-Peace	4B	37.92/46.44	42.26/52.72	59.42/70.82	74.98/84.41	18.83/27.70	84.03/89.20
E5-V	8.4B	44.51/55.01	46.30/55.93	67.57/78.65	74.58/83.58	21.60/31.06	87.76/93.25
Ours	3.4B	**80.25/87.59**	**69.18/78.10**	**84.82/91.80**	**90.31/94.30**	**42.91/58.17**	**96.42/98.51**

using a 3.4B backbone, we still outperform larger models (e.g., 8.4B E5-V) by an average of 21.3 points, underscoring that unified encoding, scaled training, and residual fusion matter more than raw scale. This integrated framework compensates for architectural constraints and offers a more efficient path to multimodal retrieval excellence.

From an aggregate perspective, our model delivers consistent gains: averaged over six benchmarks we achieve +3.99/+4.27 in Recall@1/Recall@3 versus GME, with the largest margins on chart-centric datasets (ChartQA +12.14/+11.99, PlotQA +7.48/+10.75). Text-dominant benchmarks also benefit, with ArxivQA +0.23/+0.62, DocVQA +1.49/+0.80, InfoVQA +1.42/+0.30, and SlideVQA +1.17/+1.16. The unified index further removes dual-retriever fusion, simplifying deployment while maintaining high recall.

4.4 Cross-Modal Semantic Alignment Analysis

To investigate how our scaled training strategy enhances retrieval performance, we analyzed cross-modal semantic alignment by examining query-document similarity distributions. For each dataset, we computed cosine similarities between queries and ground-truth documents categorized by modality (pure text, pure image, text-image mixed) and visualized results using box plots.

Fig. 3 demonstrates that scaled training significantly improves semantic quality across datasets, shifting all modality similarity distributions toward higher semantic spaces, enhancing both individual modality representations and coordinated cross-modal development. Consistent improvements across all six datasets (academic, industrial, infographics, scientific plots, presentations) under standard evaluation metrics demonstrate the strategy's general applicability and strong domain generalization capability, and the elevated query–document similarities directly correlate with the retrieval gains in Table 2, validating the 2:8:6 schedule as a principled way to curb cross-modal inconsistency in text-query-driven multimodal retrieval.

4.5 Ablation Study

Bimodal Fusion Mechanism. To isolate the contribution of our fusion mechanism, we compare against VisRAG [19], which uses a similar unified encoder

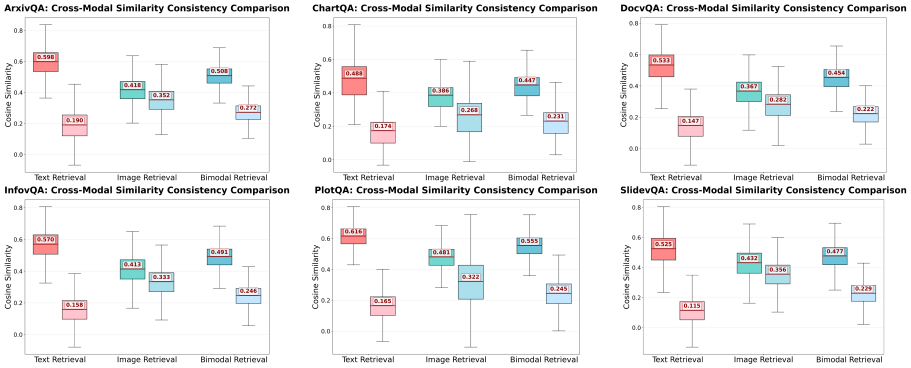

Fig. 3. Query-Document Cosine Similarity Distribution Across Modalities. Box plots show similarities for pure text, pure image, and bimodal documents. Left groups (higher medians): scaled strategy; right groups: original VisRAG.

Table 3. Ablation of Bimodal Fusion Mechanism in Recall@1/Recall@3.

Model	ArxivQA	ChartQA	DocVQA	InfoVQA	PlotQA	SlideVQA
VisRAG	35.83/45.86	42.68/51.46	60.49/71.19	73.11/79.79	16.96/23.66	80.17/87.43
Ours	**80.25/87.59**	**69.18/78.10**	**84.82/91.80**	**90.31/94.30**	**42.91/58.17**	**96.42/98.51**

Table 4. Ablation of Scaled Training Strategy in Recall@1/Recall@3.

Model	ArxivQA	ChartQA	DocVQA	InfoVQA	PlotQA	SlideVQA
1:1:1	63.92/67.65	49.37/61.09	72.74/80.56	80.68/88.05	23.81/31.75	85.61/90.41
2:8:6	**80.25/87.59**	**69.18/78.10**	**84.82/91.80**	**90.31/94.30**	**42.91/58.17**	**96.42/98.51**

architecture but employs simple averaging for bimodal fusion. Both methods use identical training configurations to ensure fair comparison, with the only difference being the fusion mechanism.

Table 3 demonstrates consistent improvements with our fusion mechanism across all datasets. The improvements are particularly substantial on visual-intensive tasks, with PlotQA showing the most dramatic gains. Even datasets with high baselines like InfoVQA and SlideVQA achieve notable improvements, overcoming performance ceiling effects. These results validate our fusion design: learnable parameters effectively capture text-visual correlations while residual connections preserve original modal information, significantly outperforming simple averaging by adapting to task-specific characteristics.

Scaled Training Strategy. To validate the effectiveness of our scaled training strategy, we compare uniform training (1:1:1 text:image:bimodal) with our proposed 2:8:6 strategy. Both configurations use identical model architecture, fusion mechanism, and training hyperparameters (batch size, learning rate, epochs), with the only difference being the training sample distribution. This controlled comparison isolates the contribution of the scaled training strategy.

Table 4 demonstrates consistent improvements with the 2:8:6 strategy across all six datasets, with PlotQA showing the most substantial gains (+19.10 Recall@1), validating our hypothesis that increased image modality exposure compensates for architectural capacity disparity. Notably, even text-rich datasets like SlideVQA still benefited from the scaled strategy (+10.81 Recall@1), indicating that enhanced visual learning strengthens cross-modal alignment even in text-heavy contexts. The improvements are consistent across all datasets, confirming that the 2:8:6 ratio effectively addresses the three imbalance factors identified in our theoretical analysis. Importantly, the 2:8:6 schedule does not increase training compute (identical batch size and steps), indicating that the gains arise purely from rebalancing modality exposure rather than additional training budget.

5 Conclusion

We propose a unified multimodal retrieval framework for RAG, addressing modality separation and inconsistency via three innovations: (1) a unified encoder mapping all modalities into a shared semantic space, (2) a learnable residual fusion mechanism adapting to task-specific cross-modal interactions, (3) a theoretically-grounded scaled training strategy compensating for architectural/optimization imbalances. Our method achieves state-of-the-art results across six benchmarks, with ablations validating each component's contribution. Our modality imbalance analysis and 2:8:6 ratio offer practical guidance for multimodal system design. Future work will extend the framework to additional modalities, explore adaptive training ratios, and integrate the retriever into end-to-end RAG pipelines for holistic evaluation.

Acknowledgements. This work was partly supported by the NSFC (62431015, 62571317, 62501387), the Fundamental Research Funds for the Central Universities, Shanghai Key Laboratory of Digital Media Processing and Transmission under Grant 22DZ2229005, and the 111 project BP0719010.

References

1. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216 (2024)
2. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. In: The Thirteenth International Conference on Learning Representations (2024)
3. Günther, M., Sturua, S., Akram, M.K., Mohr, I., Ungureanu, A., Eslami, S., Martens, S., Wang, B., Wang, N., Xiao, H.: jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. arXiv preprint arXiv:2506.18902 (2025)
4. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282 (2020)

5. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
6. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024)
7. Karpukhin, V., Oguz, B., Min, S., Lewis, P.S., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: EMNLP (1). pp. 6769–6781 (2020)
8. Li, L., Wang, Y., Xu, R., Wang, P., Feng, X., Kong, L., Liu, Q.: Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231 (2024)
9. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
10. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
11. Methani, N., Ganguly, P., Khapra, M.M., Kumar, P.: Plotqa: Reasoning over scientific plots. In: Proceedings of the ieee/cvf winter conference on applications of computer vision. pp. 1527–1536 (2020)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
13. Tanaka, R., Nishida, K., Nishida, K., Hasegawa, T., Saito, I., Saito, K.: Slidevqa: A dataset for document visual question answering on multiple images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 13636–13645 (2023)
14. Tito, R., Karatzas, D., Valveny, E.: Hierarchical multimodal transformers for multipage docvqa. Pattern Recognition **144**, 109834 (2023)
15. Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., Zhou, C.: One-peace: Exploring one general representation model toward unlimited modalities. arXiv preprint arXiv:2305.11172 (2023)
16. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1192–1200 (2020)
17. Yan, S.Q., Gu, J.C., Zhu, Y., Ling, Z.H.: Corrective retrieval augmented generation (2024)
18. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
19. Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., et al.: Visrag: Vision-based retrieval-augmented generation on multi-modality documents. arXiv preprint arXiv:2410.10594 (2024)
20. Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., Zhang, M.: Gme: Improving universal multimodal retrieval by multimodal llms. arXiv preprint arXiv:2412.16855 (2024)
21. Zhou, J., Liu, Z., Xiao, S., Zhao, B., Xiong, Y.: Vista: Visualized text embedding for universal multi-modal retrieval. arXiv preprint arXiv:2406.04292 (2024)