Multimodal RAG System: Architecture Write-Up

Embedding Strategy: Separate Spaces

The system uses two independent embedding spaces: **MiniLM-L6-v2** (384-dim) for text and **CLIP ViT-B/32** (512-dim) for images, stored in separate ChromaDB collections. A unified space (e.g., CLIP for both) was rejected because MiniLM produces substantially higher-quality text-to-text retrieval than CLIP's text encoder, which is optimized for cross-modal alignment rather than semantic textual similarity. Keeping the spaces separate allows each encoder to operate in its strength: MiniLM retrieves semantically relevant QA pairs, while CLIP's text encoder projects queries into the vision-language space for image retrieval. The tradeoff is that scores across modalities are not directly comparable: MiniLM cosine similarities typically range 0.5–0.85 for relevant results, while CLIP text-to-image similarities range 0.25–0.35. The system accounts for this by applying a confidence threshold only to text scores and relying on a group_id linkage mechanism to bridge modalities at context assembly rather than at the embedding level.

Model Choices

MiniLM-L6-v2 was selected for text encoding due to its strong performance-to-size ratio (22M parameters, fast inference), well-suited to indexing 35K+ documents. **CLIP ViT-B/32** provides a joint vision-language space out of the box for text-to-image retrieval. **Qwen3-VL-8B-Instruct** was chosen for generation as a vision-language model that runs locally on a single GPU and processes retrieved images directly alongside text evidence, avoiding external API dependencies. Rather than relying solely on text proxies (linked QA pairs) to describe images, the VLM performs genuine visual grounding: reading chart values, identifying diagram labels, and describing visual structures from pixels. This was validated empirically—on chart-reading queries, the VLM extracted specific numerical values (e.g., import volumes, prevalence percentages) directly from chart images that were absent from retrieved text. Greedy decoding (`do_sample=False`) maximizes citation determinism.

Preventing Hallucinated Citations

Citation integrity is enforced through three mechanisms. First, the **Context Builder** performs bidirectional group_id linkage: for each retrieved text, it resolves the paired image, and for each retrieved image, it resolves paired text records. This produces a valid_ids allowlist containing only document IDs from retrieval results or linked via group_id. The system prompt restricts citing to these IDs (soft constraint). Second, a post-generation **Citation Validator** extracts all bracketed citations via regex and checks each against valid_ids; any unmatched citation is flagged as `[INVALID: ...]` (hard constraint). Third, a **score threshold** on text retrieval triggers an "insufficient evidence" response when the best score falls below 0.4, preventing generation from weak evidence.

Scaling to 1M+ Documents

Vector search: ChromaDB's HNSW index works at current scale (~57K docs) but would require tuning (ef_construction, M) or migration to Milvus/Pinecone/Weaviate for 1M+ with latency constraints. **Context assembly**: the in-memory group_id lookup would be replaced by a key-value store (Redis, PostgreSQL) with indexed lookups. **Embedding throughput**: image encoding (~116 img/sec on CPU) would be parallelized across GPUs. **VLM inference**: images per query would be capped (top 3–4) with bounded resolution; Flash Attention 2 and batched generation would reduce per-query cost. Document storage would be separated from vector metadata via a dedicated store with doc_id as foreign key.

Evaluation Plan

Retrieval quality: Recall@K and MRR using AI2D and ChartQA ground-truth QA pairs, where retrieval is correct if the returned document's group_id matches the question's source group. **Citation accuracy**: ratio of valid to total citations across a held-out query set, targeting zero invalid citations. **Answer quality**: exact match and token-level F1 against ground-truth, emphasizing whether the VLM correctly extracts numerical values from charts and identifies spatial relationships in diagrams, validating genuine visual grounding rather than text-proxy reliance. End-to-end latency would also be benchmarked per query.