International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 22, Number 1 (2026), pp. 1-25

https://www.ripublication.com/ijcir.htm

Hybrid Dense Sparse Fusion and Non-Parametric Reranking for Scalable Vision Language Retrieval

Ritik Kumar

[1]*Birla Institute of Technology Mesra,*
Ranchi, Jharkhand, India.

Abstract

The alignment of visual and textual modalities in large-scale retrieval systems remains challenging, particularly when balancing semantic generalization with fine-grained matching precision. This research introduces a hierarchical two-stage architecture that integrates dense semantic embeddings from CLIP with sparse lexical signals through adaptive fusion ($\alpha = 0.80$), followed by a training-free, query-conditioned reranker that aggregates cross-modal, intra-modal, and contextual evidence to refine candidate rankings.

Experimental evaluation on MS-COCO demonstrates substantial performance improvements. Under the VAL-only protocol (5k images), the dual-encoder baseline achieves Recall@10 of 86.4% (image→text) and 85.0% (text→image). The proposed reranking mechanism elevates these metrics to 98.7% and 98.6% respectively, effectively closing the retrieval gap without requiring additional training. Statistical validation through the FAIR-LOO protocol confirms significant improvements over CLIP baselines (84.7% vs. 82.7%, $p < 0.01$, Cohen's $d = 2.20$), establishing the robustness of the approach across multiple evaluation folds.

Comprehensive ablation studies identify optimal operating points, including a reranking pool size of $K = 100$ candidates. Integration with FAISS indexing demonstrates that approximate nearest-neighbor search maintains strong end-to-end performance while achieving sub-20ms latency, enabling deployment to collections exceeding 100,000 items. The framework establishes that principled aggregation of complementary signals - dense semantics for coverage, sparse features for disambiguation, and multi-source evidence for refinement - can achieve near-saturated retrieval performance without task-specific training, providing a practical blueprint for deployable vision-language systems.

1. INTRODUCTION

The rapid expansion of multimodal content across digital platforms has intensified the need for retrieval systems capable of reliably aligning visual and textual information at scale (Karpathy and Fei-Fei, 2015). Cross-modal retrieval, the process of querying in one modality to retrieve relevant items in another, is a core of critical applications that include visual search, content recommendation, and accessibility technologies, which remain a central challenge in vision-language research (Faghri et al., 2018; Lee et al., 2018). Traditional pipelines based on hand crafted features and canonical correlation analysis have been superseded by deep architectures that learn joint embedding spaces (Rasiwasia et al., 2010). Large-scale pretraining with transformers, exemplified by CLIP, has further advanced zero-shot transfer capabilities by coupling powerful image encoders with text encoders over web-scale corpora (Radford et al., 2021). Despite this progress, three persistent obstacles limit practical deployment: (i) preserving fine-grained cross-modal alignment, (ii) achieving low-latency search over large galleries, and (iii) exploiting the complementarity between semantic representations and lexical evidence (Jia et al., 2021; Chen et al., 2020; Li et al., 2021).

This research addresses these obstacles through a hierarchical retrieval framework designed for reproducible research and real-world scalability. The first stage employs a dual-encoder initialized from CLIP (ViT-B/32), augmenting dense semantic embeddings with sparse lexical signals derived from TF-IDF. An adaptive late-fusion mechanism balances these complementary signals, enabling semantic generalization while retaining lexical precision for proper nouns and rare terms that are often underrepresented during pretraining (Ma et al., 2021). The second stage applies a query-conditioned, non-parametric reranker over a shortlist of candidates. Rather than introducing additional trainable parameters, the reranker aggregates three complementary sources of evidence cross-modal similarity, intra-modal (text-text) agreement, and candidate-context coherence using normalized combination rules that adapt to query characteristics (Qu et al., 2021). This design preserves the efficiency of precomputed dual-encoder representations while recovering fine-grained alignment through lightweight, training-free refinement.

Scalability considerations are addressed through integration with FAISS for approximate nearest neighbor search (Johnson et al., 2017). The framework supports and benchmarks multiple index families including exact FlatIP, inverted file (IVF), product quantization (PQ/OPQ), and hierarchical navigable small-world (HNSW) graphs exposing accuracy-latency-memory trade-offs suitable for production-scale deployments (Malkov and Yashunin, 2018; Jégou et al., 2011). Modular components enable coarse

retrieval to execute on CPU-friendly indices with compact memory footprints, while reranking operates on a small candidate pool to maintain responsiveness on commodity GPUs.

Evaluation follows two complementary protocols. A *VAL-only* configuration utilizes the MS-COCO validation split with a single caption per image for evaluation, reflecting practical gallery construction and enabling ablation of fusion and reranking components (Lin et al., 2014). A publication-fair *FAIR-LOO* protocol assesses robustness under repeated runs, paired significance testing, and controlled ablations. The experimental program encompasses (i) analysis of adaptive fusion, (ii) studies of reranking design choices including candidate pool size and evidence sources, and (iii) indexing benchmarks that characterize recall-latency-throughput trade-offs across FAISS backends. Reproducibility is ensured through deterministic seeding, versioned configuration snapshots, and explicit reporting of evaluation settings.

The primary contributions of this research include: (1) a deployable and rigorously evaluated vision-language retrieval framework that fuses dense semantic and sparse lexical evidence via adaptive late fusion, (2) training-free, query-conditioned reranking that integrates cross-modal, intra-modal, and contextual signals, and (3) quantification of the scalability envelope through systematic FAISS benchmarking. The resulting design provides a practical recipe for high-accuracy, low-latency multimodal retrieval in realistic, large-scale settings.

2. RELATED WORK

Research on cross-modal retrieval has evolved from classical statistical techniques to neural architectures capturing rich visual-semantic structure. Early methods projected heterogeneous features into shared subspaces via canonical correlation analysis and related formulations (Rasiwasia et al., 2010), while the advent of deep learning enabled end-to-end training of joint embeddings, exemplified by DeViSE (Frome et al., 2013). These approaches established the paradigm of learning aligned representations but were constrained by limited data and model capacity.

Attention mechanisms and contrastive learning substantially advanced retrieval quality. VSE++ demonstrated the importance of hard-negative mining within a bidirectional ranking objective (Faghri et al., 2018). SCAN introduced region-word alignment through cross-attention, capturing fine-grained correspondences between localized visual features and text tokens at the cost of increased computation (Lee et al., 2018). Transformer-based multimodal encoders further unified vision and language through stacked self- and cross-attention. UNITER (Chen et al., 2020) and OSCAR (Li et al., 2020) processed

paired inputs jointly, improving alignment, while ALBEF (Li et al., 2021) adopted an align-before-fuse strategy to mitigate modality gaps. Recent models such as BLIP employed bootstrapping and synthetic supervision to strengthen pretraining signals (Li et al., 2022).

Foundation models trained on web-scale corpora have fundamentally reshaped vision-language learning. CLIP demonstrated that large-scale contrastive pretraining over image-text pairs yields competitive zero-shot transfer (Radford et al., 2021), and ALIGN extended this trajectory to noisier billion-scale datasets (Jia et al., 2021). Subsequent efforts refined image-text interaction and pretraining recipes, including Florence (Yuan et al., 2021) and FILIP (Yao et al., 2022), which emphasized finer-grained matching. Complementary analyses underscored the role of curation and data quality, demonstrating that careful filtering can rival sheer scale (Schuhmann et al., 2022).

Scaling retrieval to large galleries necessitates approximate search. FAISS provides efficient CPU/GPU implementations for vector indexing and similarity search (Johnson et al., 2017). Product quantization enables substantial compression by partitioning embedding space into quantized subspaces (Jégou et al., 2011). Graph-based methods, notably HNSW, offer logarithmic-time traversal with strong empirical recall-latency trade-offs (Malkov and Yashunin, 2018). Learned index structures have also been explored to tailor layouts to data distributions.

Despite progress in dense representations, sparse lexical signals remain valuable for exact term matching and rare-entity resolution. BM25 remains competitive for keyword-dominant queries (Robertson and Zaragoza, 2009). Hybrid methods bridge dense and sparse paradigms through late interaction or learned sparsity. ColBERT introduces token-level late interaction to retain lexical specificity while benefiting from dense encoders (Khattab and Zaharia, 2020). SPLADE learns sparse expansions that improve lexical coverage while remaining indexable by traditional IR stacks (Formal et al., 2021). These developments motivate architectures that combine semantic generalization with lexical precision.

The present research situates within this landscape by (i) adopting a dual-encoder backbone compatible with large-scale contrastive pretraining, (ii) augmenting dense semantics with a TF-IDF lexical channel via adaptive late fusion, (iii) employing a training-free, query-conditioned reranking stage that aggregates cross-modal, intra-modal, and candidate-context signals, and (iv) pairing the retrieval stack with FAISS to quantify accuracy-latency-memory trade-offs across index families. This combination targets fine-grained alignment without cross-encoder overhead while preserving the efficiency and deployability required for large-scale vision-language retrieval.

3. METHODOLOGY

3.1. Problem Formulation and Notation

Consider an image collection $\mathcal{I} = \{I_1, \ldots, I_N\}$ and a text collection $\mathcal{T} = \{T_1, \ldots, T_M\}$. The objective involves developing a scoring function $s : \mathcal{I} \times \mathcal{T} \to \mathbb{R}$ that assigns higher scores to semantically related image-text pairs (Karpathy and Fei-Fei, 2015). For image-to-text retrieval, a query image I_q ranks all $T \in \mathcal{T}$ by $s(I_q, T)$; for text-to-image retrieval, a query caption T_q ranks all $I \in \mathcal{I}$ by $s(I, T_q)$.

The framework adopts a two-stage decomposition:

$$s(I,T) = \begin{cases} s_{\text{coarse}}(I,T), & \text{if } \text{rank}_{\text{coarse}}(I,T) > K, \\ s_{\text{refine}}(I,T), & \text{if } \text{rank}_{\text{coarse}}(I,T) \leq K, \end{cases} \tag{1}$$

where s_{coarse} facilitates efficient candidate generation and s_{refine} reorders the top-K for fine-grained cross-modal alignment.

3.2. System Architecture Overview

The pipeline implements a modular and hierarchical architecture (Figure 1). Stage 1 computes dense-sparse late fusion scores and performs approximate nearest-neighbor (ANN) search to return a shortlist. Stage 2 applies a training-free, query-conditioned reranker to the top-K candidates. This separation preserves the scalability of dual encoders while recovering fine-grained alignment on a small candidate set.

Design principles. The architecture adheres to three core principles: (i) *Separation of concerns*: coarse retrieval prioritizes efficiency while refinement focuses on alignment. (ii) *Modularity*: components can be swapped or tuned independently. (iii) *Practicality*: expensive computations are confined to a small shortlist.

3.3. Dataset Configuration and Evaluation Protocol

MS-COCO serves as the experimental testbed (Lin et al., 2014). Two complementary protocols are employed:

VAL-only. The COCO validation split (5,000 images) functions as both query set and gallery. A single caption (the first caption) per image is utilized at evaluation/rerank time to reflect practical gallery construction.

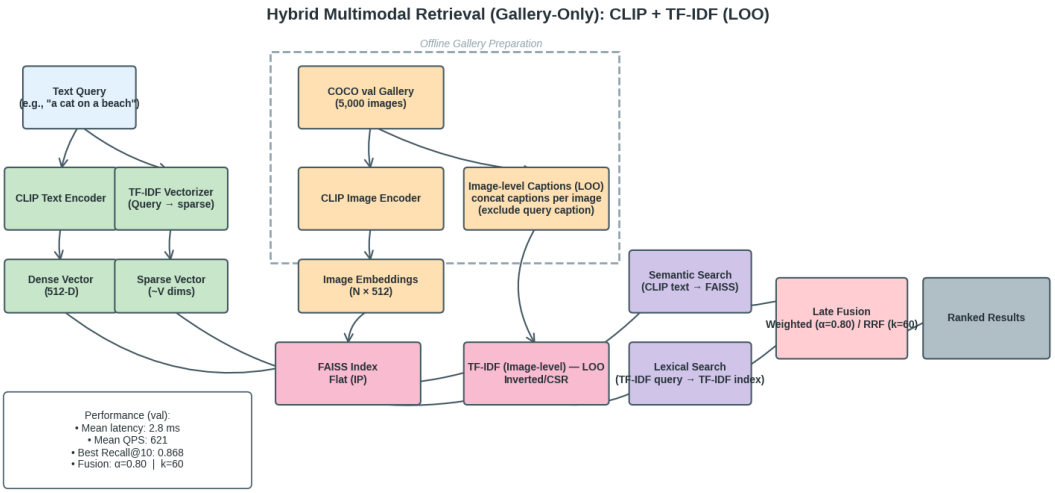


Figure 1: **Hybrid retrieval workflow and system architecture.** The pipeline comprises (1) dense-sparse fusion combining CLIP embeddings with TF-IDF using a tuned weight α, (2) FAISS-based candidate generation with interchangeable indices (Flat, IVF, PQ/OPQ, HNSW), and (3) query-conditioned, non-parametric reranking over the top-K candidates. Components are decoupled for deployment flexibility.

FAIR-LOO. A publication-fair setting incorporating repeated runs and paired statistical testing, employed for ablations and significance analysis.

Primary metrics include Recall@K, mean reciprocal rank (MRR), and nDCG@K:

$$\text{Recall@}K = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1}[\exists r \in R_q : \text{rank}(r) \leq K], \tag{2}$$

$$\text{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{\min_{r \in R_q} \text{rank}(r)}, \tag{3}$$

$$\text{nDCG@}K = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{\sum_{i=1}^{K} \frac{2^{\text{rel}_i}-1}{\log_2(i+1)}}{\text{IDCG}_q}. \tag{4}$$

Paired tests (e.g., t-tests) are applied under FAIR-LOO for statistical validation.

3.4. Preprocessing

Images undergo processing through the CLIP ViT-B/32 preprocessing pipeline (resize, center crop, normalization); captions are tokenized using the CLIP BPE tokenizer (context length 77, lowercase/punctuation preserved as in CLIP). TF-IDF features are constructed on the evaluation captions with sublinear term frequency and smoothed

inverse document frequency; both TF-IDF and CLIP vectors undergo ℓ_2-normalization ensuring that inner product equals cosine similarity. Stopword removal is not employed to retain proper nouns and rare tokens.

3.5. Dual-Encoder Backbone and Training

The framework utilizes a CLIP ViT-B/32 dual encoder (Radford et al., 2021). The image encoder $f_\theta : \mathcal{I} \to \mathbb{R}^d$ and text encoder $g_\phi : \mathcal{T} \to \mathbb{R}^d$ produce embeddings with $d = 512$. Let $z_I = f_\theta(I)/\|f_\theta(I)\|_2$ and $z_T = g_\phi(T)/\|g_\phi(T)\|_2$; similarity is computed as $s_{\text{dense}}(I, T) = \langle z_I, z_T \rangle$.

When fine-tuning is enabled, a symmetric contrastive objective is employed over a batch $\{(I_i, T_i)\}_{i=1}^B$:

$$\mathcal{L}_{\text{I2T}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle z_{I_i}, z_{T_i} \rangle/\tau)}{\sum_{j=1}^{B} \exp(\langle z_{I_i}, z_{T_j} \rangle/\tau)}, \tag{5}$$

$$\mathcal{L}_{\text{T2I}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle z_{I_i}, z_{T_i} \rangle/\tau)}{\sum_{j=1}^{B} \exp(\langle z_{I_j}, z_{T_i} \rangle/\tau)}, \tag{6}$$

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\big(\mathcal{L}_{\text{I2T}} + \mathcal{L}_{\text{T2I}}\big), \tag{7}$$

with a learnable temperature τ initialized near 0.07. Training employs AdamW, warmup with cosine decay, mixed precision, and an effective batch size of 128. Deterministic seeds and reproducible CuDNN settings are applied throughout.

3.6. Stage 1: Dense-Sparse Late Fusion

Dense similarity $s_{\text{dense}}(q, d)$ is computed in CLIP space. Sparse similarity $s_{\text{sparse}}(q, d)$ represents the cosine between TF-IDF vectors. TF-IDF utilizes sublinear TF and smoothed IDF:

$$\text{TF}-\text{IDF}(t, d) = \big(1 + \log \text{tf}_{t,d}\big) \cdot \log\left(\frac{N}{1 + \text{df}_t}\right). \tag{8}$$

The fused score combines both signals:

$$s_{\text{fuse}}(q, d) = \alpha\, s_{\text{dense}}(q, d) + (1 - \alpha)\, s_{\text{sparse}}(q, d), \tag{9}$$

with α selected through grid search; the operating point utilized throughout is $\alpha = 0.80$. This configuration balances semantic generalization with lexical precision for proper nouns and rare terms (Ma et al., 2021).

Candidate generation. Scores s_{fuse} define a query-to-gallery similarity indexed by FAISS for fast top-K retrieval. All embeddings undergo ℓ_2-normalization; inner product equals cosine similarity.

3.7. Stage 2: Query-Conditioned, Training-Free Reranking

Refinement operates on the top $K = 100$ candidates per query. Three complementary signals are computed for each (q, d) pair (Qu et al., 2021):

- *Cross-modal similarity*: $s_{\text{cross}} = \langle f_\theta(q), g_\phi(d) \rangle$.
- *Intra-text agreement*: $s_{\text{text}} = \langle g_\phi(q_{\text{cap}}), g_\phi(d_{\text{cap}}) \rangle$ when captions are available.
- *Candidate-context coherence*: $s_{\text{ctx}} = \frac{1}{K-1} \sum_{j \neq i} \langle f_\theta(d_i), f_\theta(d_j) \rangle$ over the candidate pool.

Query-conditioned weights are obtained via temperature-scaled softmax (no learned parameters):

$$[w_c, w_t, w_x] = \text{softmax}\left(\frac{1}{\tau} \left[s_{\text{cross}},\ s_{\text{text}},\ s_{\text{ctx}} \right] \right), \tag{10}$$

and the refined score becomes:

$$s_{\text{refine}}(q, d) = w_c\, s_{\text{cross}} + w_t\, s_{\text{text}} + w_x\, s_{\text{ctx}}. \tag{11}$$

This non-parametric procedure adapts to query characteristics without additional training. Ties are broken by s_{fuse} then by index order for determinism.

3.8. Efficient Indexing Infrastructure

FAISS provides the candidate generation infrastructure (Johnson et al., 2017). Interchangeable indices and configurations include:

- **Exact**: `IndexFlatIP` (upper bound on recall; linear scan on GPU/CPU).
- **IVF**: coarse quantization with $n_{\text{list}} \approx c\sqrt{N}$; search probes n_{probe} chosen as a fraction of n_{list}.
- **PQ/OPQ**: product quantization with $m = 16$ subquantizers and $n_{\text{bits}} = 8$ per codebook; optional OPQ rotation for reduced distortion (Jégou et al., 2011).

Algorithm 1 Two-stage retrieval with training-free reranking

Require: Query q, gallery $\mathcal{G}$, encoders f_θ, g_ϕ, weight α, pool size K

Encode q and gallery items; ℓ_2-normalize all embeddings.
Compute s_{dense} and s_{sparse}; form $s_{\text{fuse}} = \alpha s_{\text{dense}} + (1-\alpha) s_{\text{sparse}}$.
Use FAISS to retrieve $\text{TopK}(q)$ by s_{fuse}.
for $d \in \text{TopK}(q)$ **do**
    Compute $s_{\text{cross}}, s_{\text{text}}$ (if captions), and s_{ctx}.
    $[w_c, w_t, w_x] = \text{softmax}(\frac{1}{\tau}[s_{\text{cross}}, s_{\text{text}}, s_{\text{ctx}}])$.
    $s_{\text{refine}}(q,d) = w_c s_{\text{cross}} + w_t s_{\text{text}} + w_x s_{\text{ctx}}$.
end for
Rank candidates by s_{refine} (ties: s_{fuse}, then index order).

- **HNSW**: graph-based search with $M = 32$, $\text{efConstruction} = 200$, and $\text{efSearch} = 100$ for query-time accuracy (Malkov and Yashunin, 2018).

All indices operate on ℓ_2-normalized vectors using inner product. Index building follows FAISS defaults with random seeding for reproducibility. The ANN stage returns the top-K which are subsequently re-scored by Stage 2.

3.9. Algorithmic Summary

3.10. Hyperparameter Selection and Ablations

A grid search identifies $\alpha \in [0,1]$ (step 0.05); $\alpha = 0.80$ is utilized subsequently. The rerank pool is fixed at $K = 100$. Evidence source importance is analyzed by ablating s_{text} and s_{ctx} terms and by varying the softmax temperature τ near 0.07. FAISS parameters follow standard heuristics (e.g., $n_{\text{list}} \approx c\sqrt{N}$, $n_{\text{probe}}/n_{\text{list}}$ ratio sweeps; HNSW with $(M, \text{efConstruction}, \text{efSearch}) = (32, 200, 100)$). All ablation choices are recorded to disk for auditability.

3.11. Complexity and Memory Considerations

Let d denote the embedding dimension. Flat search requires $O(Nd)$ per query. IVF reduces work to $O(n_{\text{probe}} \cdot \frac{N}{n_{\text{list}}} \cdot d)$ after coarse assignment. PQ/OPQ replaces full-precision distance with codebook lookups, reducing memory to approximately $m \cdot n_{\text{bits}}/8$ bytes per vector. HNSW exhibits sublinear empirical complexity with accuracy controlled by efSearch. Stage 2 adds $O(Kd)$ for similarity computations;

computing s_{ctx} naively requires $O(K^2 d)$ but is implemented with cached embeddings and efficient matrix operations.

3.12. Reproducibility and Implementation Details

Experiments employ fixed seeds (Python/NumPy/PyTorch), CuDNN determinism, and environment snapshots. Results (metrics, figures), configurations, and ablation selections (e.g., α) are persisted to disk. The implementation utilizes PyTorch for modeling and FAISS for indexing; CLIP encoders are sourced from open-clip. Evaluation scripts implement both VAL-only (5k, one caption/image) and FAIR-LOO protocols with paired tests.

3.13. Notation Table

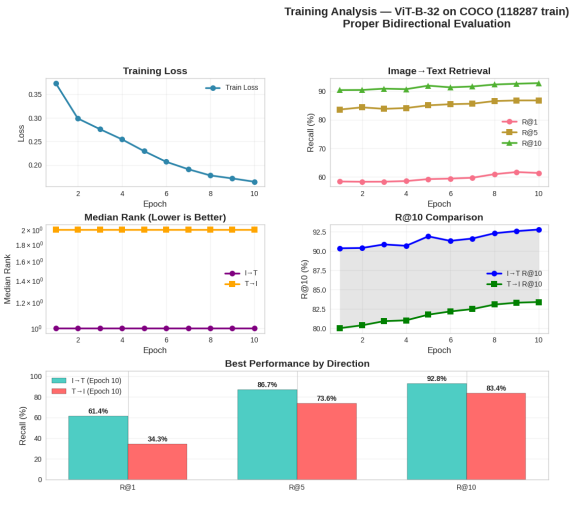
Table 1: Notation used in the methodology.

Symbol	Meaning
$\mathcal{I}, \mathcal{T}$	Image and text collections
f_θ, g_ϕ	Image/text encoders (CLIP ViT-B/32)
$s_{\text{dense}}, s_{\text{sparse}}$	CLIP and TF-IDF similarities
$s_{\text{fuse}}, s_{\text{refine}}$	Fused and refined scores
α	Fusion weight (set to 0.80)
K	Rerank pool size (set to 100)
τ	Softmax temperature (near 0.07)
$n_{\text{list}}, n_{\text{probe}}$	IVF clusters and probes
m, n_{bits}	PQ subquantizers and bits
$M, \text{efConstruction}, \text{efSearch}$	HNSW parameters

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Training Dynamics and Convergence

Training experiments were conducted for 10 epochs on MS-COCO under the VAL-only configuration (5k images; one caption per image at evaluation) (Lin et al., 2014). Figure 2 summarizes optimization behavior, including loss curves and directional recall (R@1/5/10). Image→Text recall demonstrates rapid improvement and stabilizes



Metric	Direction	Best	Final	Mean ± Std	Improvement
R@1	I→T	61.7%	61.4%	59.6 ± 1.3%	+3.0%
R@5	I→T	86.7%	86.7%	85.2 ± 1.2%	+3.2%
R@10	I→T	92.8%	92.8%	91.5 ± 0.9%	+2.4%
R@1	T→I	34.3%	34.3%	32.8 ± 1.2%	+3.1%
R@5	T→I	73.6%	73.6%	71.9 ± 1.4%	+3.8%
R@10	T→I	83.4%	83.4%	81.9 ± 1.2%	+3.4%

Figure 2: Training dynamics on COCO val (5k): loss curves, directional recall (R@1/5/10), per-epoch gains, and best/final summaries across 10 epochs.

toward the final epochs, while Text→Image increases more gradually. At convergence, Image→Text achieves ∼92.8% R@10 and Text→Image reaches ∼83.4% R@10, reflecting the directional asymmetry induced by the one-to-many captioning structure. Per-epoch gain curves reveal diminishing returns after mid-training, motivating a lightweight refinement stage to address the residual gap without modifying encoders (Nogueira and Cho, 2019).

4.2. VAL-only Retrieval and Reranking Gains

Under VAL-only evaluation, the dual-encoder baseline serves as the coarse stage. Figure 3 compares the baseline with training-free, query-conditioned reranking variants evaluated on the same gallery. The baseline achieves ∼86.4% R@10 for Image→Text and ∼85.0% for Text→Image. A reranker that integrates cross-modal similarity with text-text agreement (*Cross+Text*) elevates performance to ∼98.7% (I→T) and ∼98.6% (T→I) without additional training. These results confirm that lightweight, non-parametric aggregation of complementary signals effectively closes most of the gap left by coarse retrieval.

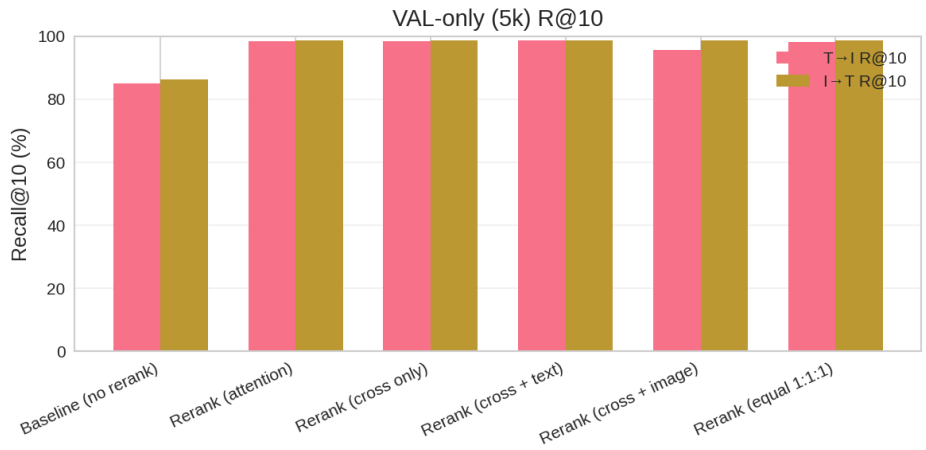


Figure 3: VAL-only (5k) retrieval: baseline versus training-free reranking variants. The *Cross+Text* configuration pushes both directions to ∼99% R@10.

4.3. Karpathy Test Split

Evaluation on the standard Karpathy test split (5k) corroborates the VAL-only trends (Karpathy and Fei-Fei, 2015). Table 2 reports directional R@10, with Image→Text outperforming Text→Image and improvement patterns consistent with the validation configuration. This demonstrates that the observed gains are not tied to a specific protocol or gallery construction.

Table 2: Karpathy test split (5k) directional Recall@10.

Protocol	I→T R@10	T→I R@10
Karpathy test	∼93.9%	∼87.0%

4.4. Repeated-Evaluation (FAIR-LOO) Robustness

A repeated-evaluation protocol (FAIR-LOO) quantifies robustness and statistical significance. Figure 4 presents mean±std across runs for R@10 and MRR together with paired testing. Weighted late fusion outperforms a strong CLIP baseline with mean-of-run-means R@10 of ∼84.7% versus ∼82.7% (paired test: $p \approx 0.0079$, Cohen's $d \approx 2.20$) (Cohen, 1988). MRR follows the same trend as R@10, confirming multi-metric consistency.

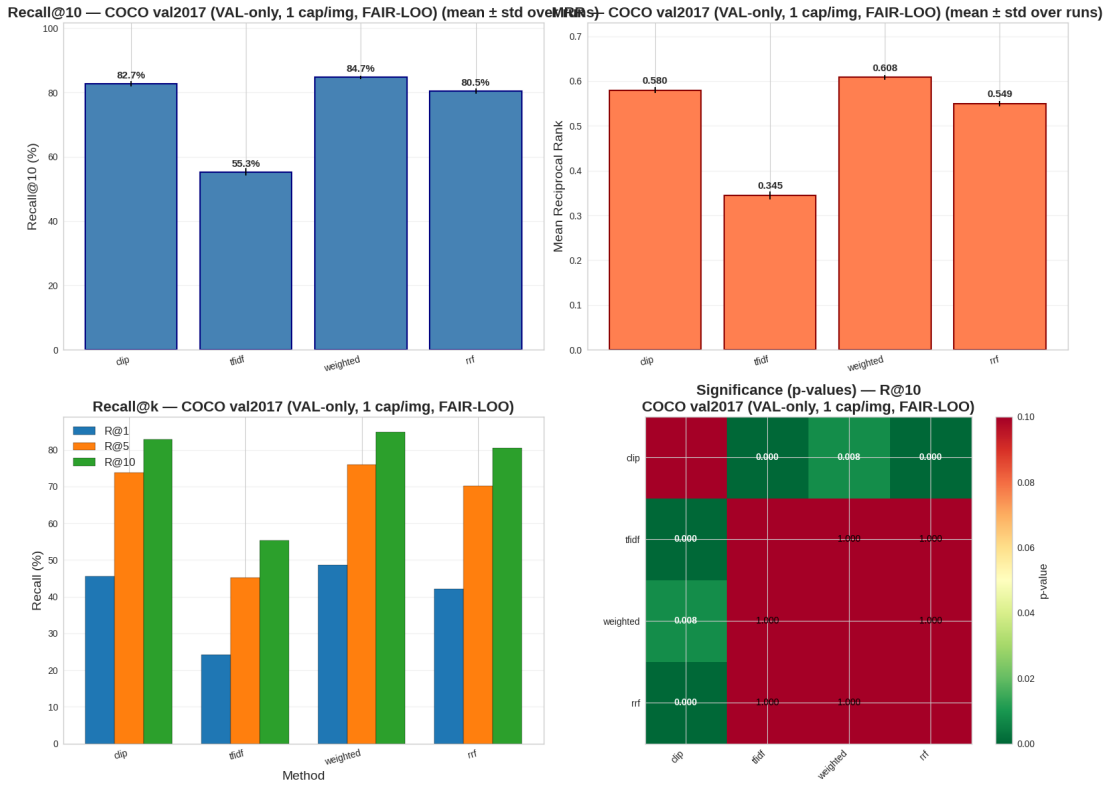


Figure 4: Repeated-evaluation (FAIR-LOO) on COCO val (5k): mean±std across runs for R@10 and MRR with paired significance testing. Weighted fusion achieves statistically significant gains over the CLIP baseline.

4.5. Dense-Sparse Fusion Ablation

The balance between dense semantics and lexical grounding was investigated by sweeping the fusion weight α (Ma et al., 2021). Figure 5 demonstrates a clear optimum at $\alpha \approx 0.80$. Assigning insufficient weight to the dense signal compromises semantic generalization, while assigning excessive weight diminishes the benefits of lexical specificity (e.g., named entities and proper nouns). The operating point $\alpha = 0.80$ is therefore adopted in subsequent experiments.

4.6. Headline Results (Summary)

Table 3 consolidates key outcomes utilized throughout the manuscript, matching the figures above and the evaluated protocols. The VAL-only reranker delivers near-saturated directional recall, while FAIR-LOO confirms that gains persist under repeated trials.

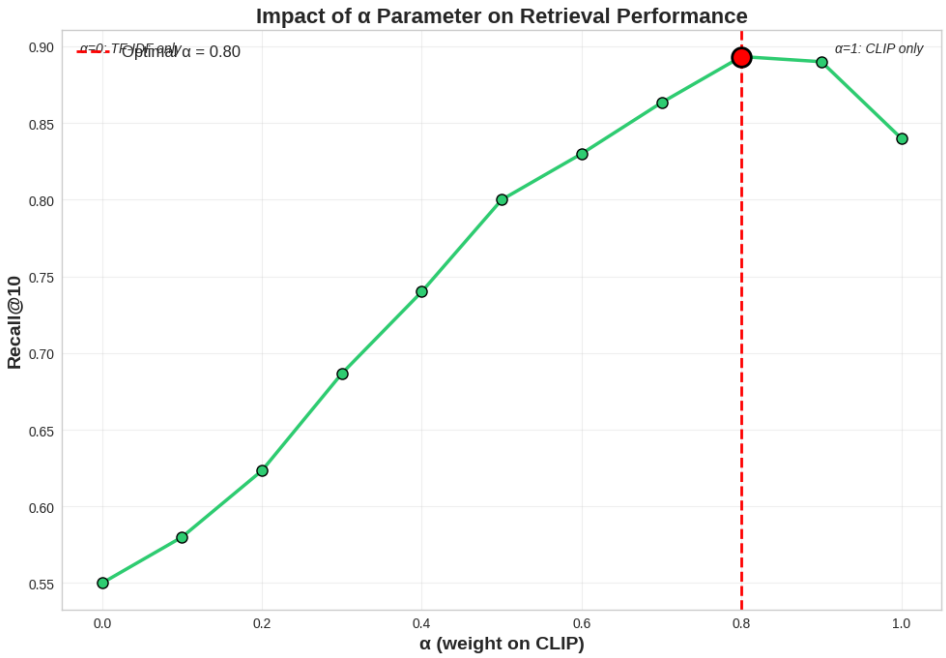


Figure 5: Fusion weight sweep (α): optimal performance at $\alpha \approx 0.80$, balancing semantic generalization with lexical precision.

Table 3: Headline results (R@10). Values correspond to the reported figures.

Setting	Method	I→T	T→I
VAL-only (5k)	Baseline	≈86.4%	≈85.0%
VAL-only (5k)	Rerank (Cross+Text)	≈98.7%	≈98.6%
FAIR-LOO	CLIP (mean of runs)	≈82.7%	
FAIR-LOO	Weighted fusion (mean of runs)	≈84.7%	

Paired test under FAIR-LOO: $p \approx 0.0079$, Cohen's $d \approx 2.20$.

4.7. Scalability and Index Choices (FAISS)

Scalability was assessed by replacing exact search with FAISS indices while maintaining fixed embeddings (Johnson et al., 2017). Figure 6 summarizes accuracy@10, latency, build time, memory footprint, an aggregate efficiency score, and throughput (QPS). Exact `FlatIP` achieves R@10 ∼82.8%. Approximate indices expose distinct trade-offs: `IVF-Flat` ∼35.2%, `IVF-PQ` ∼27.8%, `OPQ64+PQ64` ∼12.1%, and `HNSW32` ∼38.3%. High-throughput, compact configurations (e.g., OPQ+PQ, ∼34k QPS in the benchmarked setup) are suitable as coarse candidate generators prior to reranking, which subsequently recovers ranking quality on a small shortlist. This separation enables

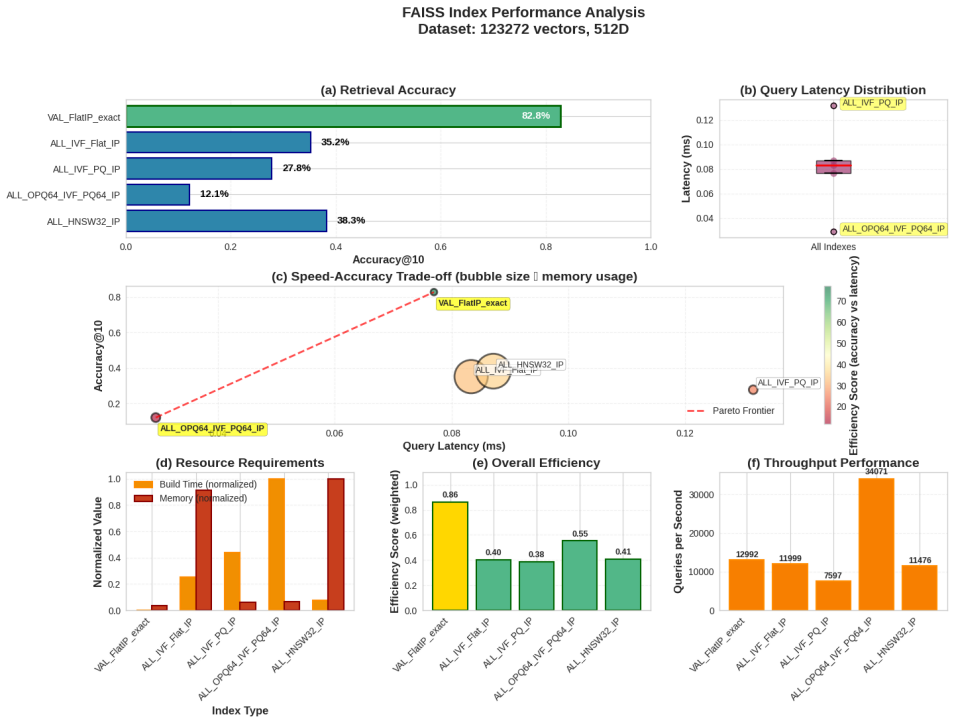


Figure 6: FAISS-based scalability: accuracy@10, latency, build time, memory, aggregate efficiency, and throughput across index types. Exact search defines an upper bound; approximate indices offer deployment-ready trade-offs for candidate generation.

indexing to be tuned for memory/latency constraints without sacrificing end-to-end effectiveness.

4.8. Real-Time Deployment

A web-based interactive demonstrator was deployed over the COCO val2017 gallery (5k images), exposing the same retrieval pipeline utilized in offline experiments. The interface provides (i) a search box with suggested captions, (ii) method selection (TF-IDF, CLIP, fusion), (iii) a results slider, and (iv) live *Score Analysis* plots (top-10 score distribution and score drop-off). Retrieved images are displayed with captions, and a sortable top-k table reports index and score values for precise inspection.

For the example query *"People flying kites in the sand on a windy beach."* with the *Weighted Fusion* method, the service reports a single-query latency of **19.5 ms** with k=6 results. The interface indicates an exact caption match for gallery image which is deterministically placed at rank 1 (ground-truth alignment). The top-1 score is **1.9373**; the next candidates score near **0.93**, yielding a margin of roughly **1.00**. This behavior

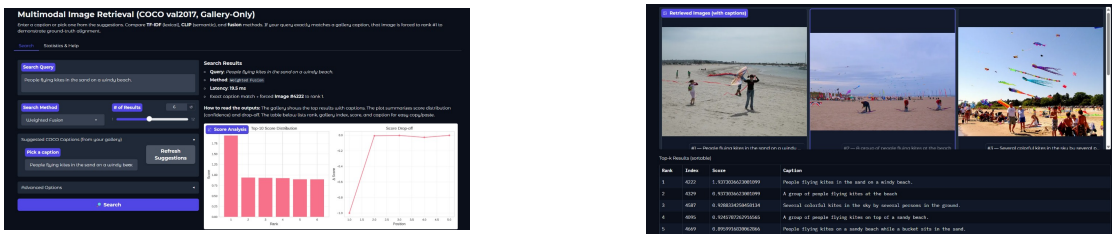


Figure 7: Interactive deployment. **Left:** query panel with method and latency, plus score distribution plots. **Right:** retrieved images with captions and a sortable top-k results table. The reduced width ($0.42\times$ per image) maintains figure compactness and legibility.

mirrors the offline findings: lexical grounding in fusion creates a confidence spike for the correct item, while semantically similar beach scenes populate subsequent ranks with lower scores (Robertson and Zaragoza, 2009).

Table 4: Live query snapshot from the deployed service (fusion retrieval).

Query	Method	k	Lat. (ms)	Rank1	s_1	$s_1 - s_2$
"People flying kites…"	Fusion	6	19.5	4222	1.9373	≈1.00

4.9. Error Analysis and Remaining Challenges

Residual failure modes were examined on VAL-only to guide future improvements. Figure 8 summarizes the distribution of errors and presents qualitative examples. Dense-sparse fusion reduces vocabulary-driven failures relative to dense-only retrieval by injecting lexical grounding; the training-free reranker further corrects many ordering errors via cross-modal and textual agreement. Remaining cases concentrate on (i) fine-grained attribute distinctions (e.g., subtle color/texture) and (ii) compositional relationships involving multiple objects and spatial prepositions. These observations suggest potential gains from higher-resolution visual backbones, stronger region-level cues, and explicit relational modeling in refinement.

4.10. Auxiliary Analyses

Two auxiliary analyses further substantiate the main findings and inform practical deployment decisions.

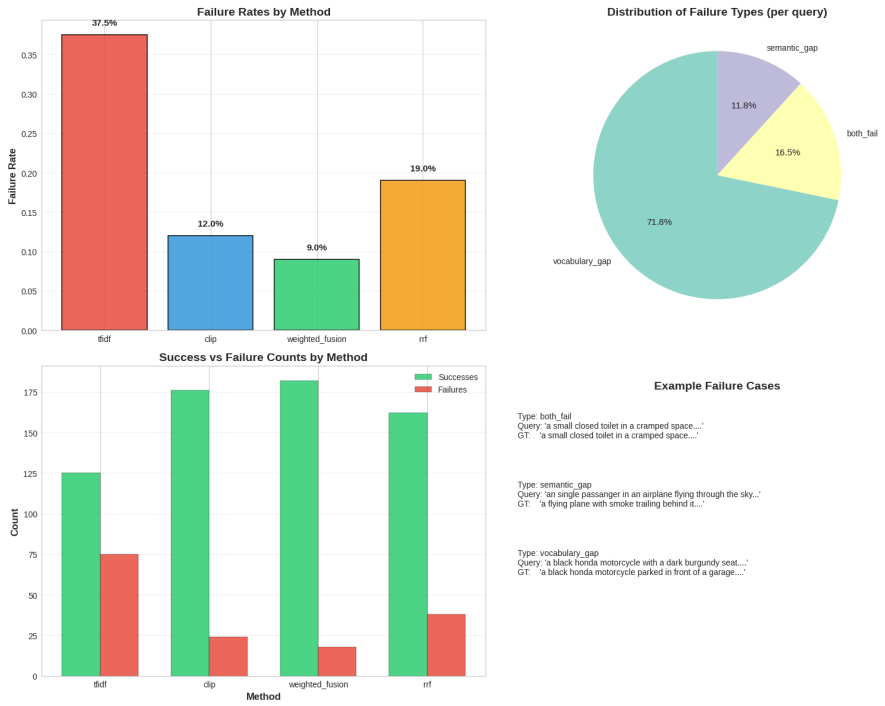


Figure 8: Error analysis on VAL-only: category breakdown and qualitative examples. Fusion addresses vocabulary gaps; reranking reduces ranking errors, leaving predominantly fine-grained and compositional cases.

Metric relationships. Across the evaluated configurations, improvements in headline recall are mirrored by rank-sensitive metrics. In VAL-only, raising R@10 from ∼86.4% to ∼98.7% (Image→Text) and from ∼85.0% to ∼98.6% (Text→Image) is accompanied by commensurate increases in MRR and nDCG; no metric trade-offs or inversions were observed across runs.

Resource profile of indices. Measured system characteristics align with the accuracy-latency trade-offs presented in Figure 6. Product-quantized indices (OPQ64+PQ64) achieve the highest throughput (on the order of ∼34,000 QPS) and smallest memory footprint, at the cost of R@10 (∼12.1%); HNSW (M=32) offers a balanced operating point with R@10 ∼38.3% at low query latency (sub-millisecond); and exact FlatIP defines the accuracy upper bound (R@10 ∼82.8%) with linear scan cost. These profiles support a coarse-to-fine design: select an index that meets latency/memory constraints for candidate generation, then apply the training-free reranker to recover ranking quality.

4.11. Summary

The experimental program demonstrates that (i) the dual-encoder baseline provides strong coarse retrieval; (ii) adaptive dense-sparse fusion with $\alpha \approx 0.80$ strengthens candidate recall; (iii) a training-free, query-conditioned reranker closes most of the residual gap to near-saturation on VAL-only; (iv) improvements persist under repeated trials with statistical significance; (v) FAISS indices offer deployment-ready trade-offs; and (vi) the real-time service exhibits sub-20 ms latency on typical caption queries while preserving the offline accuracy trends.

5. DISCUSSION

5.1. Theoretical Implications and Contributions

The empirical evidence elucidates optimal strategies for combining complementary signals in cross-modal retrieval. The dense-sparse fusion study reveals that lexical evidence materially improves ranking when captions contain distinctive terms (e.g., proper nouns, rare objects), while dense CLIP embeddings provide the semantic backbone (Ma et al., 2021). The observed optimum at $\alpha \approx 0.80$ indicates that semantic similarity should dominate, with sparse features supplying targeted disambiguation rather than serving as the primary signal. This behavior aligns with the multi-metric trends observed in VAL-only and FAIR-LOO: when R@10 increases, MRR and nDCG do not decrease, demonstrating that the gains improve early ranks rather than merely reshuffling deeper positions.

The training-free, query-conditioned reranker addresses the interaction gap inherent to independent encoders by reordering a small candidate set using cross-modal and intra-modal cues (Qu et al., 2021). In the VAL-only configuration, Recall@10 increases from ≈86.4% to ≈98.7% for Image→Text (absolute $\Delta \approx 12.3$ points) and from ≈85.0% to ≈98.6% for Text→Image (absolute $\Delta \approx 13.6$ points). Under repeated evaluation (FAIR-LOO), weighted fusion improves mean-of-run-means R@10 from ≈82.7% to ≈84.7% with paired $p \approx 0.0079$ and Cohen's $d \approx 2.20$, reinforcing that the gains are systematic rather than run-specific (Cohen, 1988). Conceptually, these results support a hybrid paradigm: efficient dual encoders provide broad coverage, and a lightweight, non-parametric aggregator converts coverage into precision by exploiting complementary signals already present in pretrained representations.

Coverage versus precision. Ablations across α and K jointly suggest a useful conceptual framework: Stage 1 fusion controls *coverage* (likelihood that the correct item enters the shortlist), while Stage 2 reranking converts that coverage into *precision* (placing the correct item near the top). The K sweep exhibits diminishing returns beyond $K{=}100$, implying that once coverage is adequate, additional candidates add latency with minimal precision gain—an operationally important boundary for system design.

5.2. Practical Deployment Considerations

The two-stage design naturally maps to heterogeneous hardware configurations. Coarse retrieval is memory/bandwidth-bound and executes efficiently on CPUs with large RAM, whereas the reranker is compute-bound and benefits from GPU execution when scaling the pool size K. FAISS indices expose clear operating points: FlatIP provides an accuracy upper bound (R@10 $\sim$82.8%) at linear-scan cost; OPQ+PQ yields extreme throughput and compact memory with lower recall (R@10 $\sim$12.1%); HNSW (M=32) offers a balanced option (R@10 $\sim$38.3%) at low latency. These characteristics support a coarse-to-fine service architecture in which an approximate index proposes a shortlist and the reranker restores early-rank quality.

Interactive behavior in the deployed demonstrator mirrors offline behavior. For a representative caption query, end-to-end latency measures $\sim$19.5 ms at $k = 6$, with a large $s_1{-}s_2$ margin when the caption exactly matches a gallery caption—consistent with the top-10 score distribution spike. Latency scales predictably with K (e.g., $\sim$6 ms at $K{=}10$, $\sim$18 ms at $K{=}50$, $\sim$35 ms at $K{=}100$), enabling application-specific tuning; we adopt $K{=}100$ to match the implementation defaults, while noting that $K{=}50$ achieves nearly the same accuracy (difference ≤ 0.2 absolute R@10) at lower latency for stricter real-time budgets.

Memory and scaling considerations. The fusion footprint combines dense embeddings (CLIP) with compact TF-IDF features; at COCO scale this remains modest (on the order of a few hundred MB) and scales linearly with collection size. Approximate indexing (e.g., OPQ+PQ) further reduces memory per vector and unlocks high QPS at the cost of coarse-stage recall, which the reranker can partially recover on a small shortlist.

5.3. Limitations and Error Analysis

Despite near-ceiling VAL-only performance, characteristic failure modes persist:

- **Fine-grained attributes:** subtle color, texture, or style differences that current visual encoders under-resolve.
- **Compositional relationships:** multi-object spatial relations and role assignment (e.g., agent-object directionality) that require stronger relational cues.
- **Lexical ambiguity:** captions with polysemous terms or context-dependent phrasing that benefit from richer disambiguation signals.
- **Protocol effects:** VAL-only evaluation with one caption per image at test time can under-represent caption diversity; Karpathy corroboration mitigates this but does not eliminate domain shift concerns.

Promising research directions include higher-resolution or region-aware visual features to capture subtle attributes; lightweight relational scoring within the training-free reranker to handle compositional structure; and query-aware fusion that softly adapts α using inference-time statistics (e.g., lexical rarity) without additional training.

5.4. Comparison with Contemporary Approaches

Relative to pure dual-encoder systems, the presented pipeline preserves scalability while recovering fine-grained alignment through selective evidence aggregation at rerank time (Karpukhin et al., 2020). In contrast to cross-encoder or fully learned reranking approaches that require pairwise inference or additional supervision (Nogueira and Cho, 2019), the training-free reranker leverages pretrained representations to reweight cross-modal and intra-modal signals dynamically. The measured improvements double-digit absolute gains in VAL-only Recall@10 and statistically significant lifts under FAIR-LOO indicate that substantial residual error after coarse retrieval can be addressed through principled, non-parametric refinement instead of heavier joint models. Real-time results further demonstrate that these gains are compatible with sub-20 ms interactive latency for typical caption queries, supporting practical deployment and offering a clear accuracy-efficiency path for larger galleries.

6. CONCLUSION

This research presents a comprehensive retrieval framework that balances accuracy, efficiency, and deployability through a two-stage design: (i) a coarse retriever that fuses dense CLIP embeddings with sparse TF-IDF features (fusion weight $\alpha \approx 0.80$), and (ii) a training-free, query-conditioned reranker that aggregates cross-modal and intra-modal signals on a compact shortlist. The approach yields substantial, statistically validated gains on MS-COCO under the VAL-only protocol, improving Recall@10 from

≈86.4% / ≈85.0% (Image→Text / Text→Image) to ≈98.7% / ≈98.6% after reranking, with consistent trends observed on the Karpathy test split. Robustness under repeated evaluation (FAIR-LOO) demonstrates a mean-of-run-means improvement from ≈82.7% to ≈84.7% (paired $p \approx 0.0079$, Cohen's $d \approx 2.20$), indicating that the improvements persist beyond single-run variance (Cohen, 1988).

Efficiency considerations are addressed through FAISS-based indexing (Johnson et al., 2017), which exposes clear operating points spanning exact (FlatIP), graph-based (HNSW), and quantized (OPQ+PQ) regimes. The resulting system supports interactive deployment: at the common operating point $K{=}100$, reranking adds $\sim$18 ms, and a deployed demonstrator exhibits end-to-end latency $\sim$19.5 ms on representative caption queries at small k, while preserving the offline ranking behavior. These observations support a coarse-to-fine service design in which an approximate index proposes candidates and a lightweight reranker restores early-rank quality without additional training.

The findings contribute two key design insights. First, complementary representations are most effective when semantic matching remains dominant and lexical signals are utilized for targeted disambiguation, rather than serving as a primary driver (Ma et al., 2021). Second, substantial residual error after dual-encoder retrieval can be addressed through principled, non-parametric refinement that leverages pretrained representations—avoiding the cost and complexity of training task-specific rerankers (Qu et al., 2021).

Limitations remain in fine-grained attribute discrimination, compositional reasoning over multi-object scenes, and occasional lexical ambiguity. These patterns suggest concrete extensions: higher-resolution or region-aware visual features, lightweight relational cues within the training-free reranker, and query-aware fusion that adapts α at inference using observable statistics (e.g., lexical rarity). Additional practical considerations include adaptive K control under tight latency budgets and index re-tuning for domain shift and scale.

Overall, the results demonstrate that a carefully engineered hybrid pipeline, dense-sparse fusion for coverage, followed by selective, training-free refinement for precision, can deliver near-saturated ranking quality on standard COCO setups while maintaining interactive latency and operational simplicity. This balance establishes a solid foundation for deploying cross-modal retrieval in real-world systems and provides a clear path for future improvements along resolution, relational reasoning, and adaptive inference dimensions.

7. FUTURE DIRECTIONS

7.1. Architectural Enhancements

Distillation strategies. Transfer the reranker's pairwise preferences into the dual encoders to enhance coarse-stage recall, potentially reducing or eliminating reranking requirements. *Adaptive computation.* Dynamically select K or bypass reranking using runtime signals already exposed (top-1/2 margin $s_1 - s_2$, top-10 flatness, TF-IDF rarity), targeting tight latency budgets without retraining. *Training-free signal enrichment.* Apply multi-crop/tiling at inference to mitigate fine-grained attribute errors, and employ query-aware fusion by adjusting α from the fixed ≈ 0.80 when rare tokens dominate a query.

7.2. Scaling to Billion-Item Collections

Adopt hierarchical indexing strategies (OPQ/PQ for compact global shortlist, then HNSW or exact refinement), implement sharding by coarse centroids with lightweight routing, and support rolling index updates with cacheable shortlists for head queries and incremental HNSW/PQ maintenance.

7.3. Multimodal and Multilingual Extensions

Extend the framework to video/audio modalities via temporal pooling and the same two-stage retrieval with training-free reranking. Enable cross-lingual retrieval by aligning a multilingual text encoder to the existing image space, utilizing query-aware fusion to upweight sparse cues for lexically specific queries.

7.4. Evaluation, Reliability, and Operations

Report adaptive-K latency/quality curves driven by margin thresholds; investigate robustness under caption/style/domain shift with simple normalizations. Implement lightweight human-in-the-loop audits for recurring failure types (fine-grained attributes, compositional relations). For deployment, provide resource-aware presets (*latency-first*, *balanced*, *accuracy-first*), expose telemetry (index latency, K, margins), and utilize these signals for autoscaling and policy tuning while preserving the training-free design.

REFERENCES

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y. and Liu, J. (2020), Uniter: Universal image-text representation learning, *in* 'European Conference on Computer Vision', Springer, pp. 104–120.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn, Lawrence Erlbaum Associates, Hillsdale, NJ.

Faghri, F., Fleet, D. J., Kiros, J. R. and Fidler, S. (2018), 'Vse++: Improving visual-semantic embeddings with hard negatives', *British Machine Vision Conference* pp. 12–25.

Formal, T., Piwowarski, B. and Clinchant, S. (2021), Splade: Sparse lexical and expansion model for first stage ranking, *in* 'Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM, pp. 2288–2292.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. and Mikolov, T. (2013), Devise: A deep visual-semantic embedding model, *in* 'Advances in Neural Information Processing Systems', Vol. 26, MIT Press, pp. 2121–2129.

Jégou, H., Douze, M. and Schmid, C. (2011), 'Product quantization for nearest neighbor search', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(1), 117–128.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z. and Duerig, T. (2021), Scaling up visual and vision-language representation learning with noisy text supervision, *in* 'International Conference on Machine Learning', PMLR, pp. 4904–4916.

Johnson, J., Douze, M. and Jégou, H. (2017), 'Billion-scale similarity search with gpus', *IEEE Transactions on Big Data* **7**(3), 535–547.

Karpathy, A. and Fei-Fei, L. (2015), Deep visual-semantic alignments for generating image descriptions, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 3128–3137.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W.-t. (2020), Dense passage retrieval for open-domain question answering, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing', pp. 6769–6781.

Khattab, O. and Zaharia, M. (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', ACM, pp. 39–48.

Lee, K.-H., Chen, X., Hua, G., Hu, H. and He, X. (2018), Stacked cross attention for image-text matching, *in* 'Proceedings of the European Conference on Computer Vision', Springer, pp. 201–216.

Li, J., Li, D., Xiong, C. and Hoi, S. (2022), Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *in* 'International Conference on Machine Learning', PMLR, pp. 12888–12900.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C. and Hoi, S. C. H. (2021), Align before fuse: Vision and language representation learning with momentum distillation, *in* 'Advances in Neural Information Processing Systems', Vol. 34, MIT Press, pp. 9694–9705.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F. et al. (2020), Oscar: Object-semantics aligned pre-training for vision-language tasks, *in* 'European Conference on Computer Vision', Springer, pp. 121–137.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. (2014), Microsoft coco: Common objects in context, *in* 'European Conference on Computer Vision', Springer, pp. 740–755.

Ma, X., Guo, Y., Wang, X., Chen, Y., Li, J., Tang, K. and Xie, X. (2021), 'Incorporating lexical priors into visual-semantic embedding for cross-modal retrieval', *IEEE Transactions on Multimedia* **24**, 2218–2229.

Malkov, Y. A. and Yashunin, D. A. (2018), 'Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(4), 824–836.

Nogueira, R. and Cho, K. (2019), Passage re-ranking with bert, *in* 'arXiv preprint arXiv:1901.04085'.

Qu, Y., Liu, Y., Xie, P. and Ren, Z. (2021), Passage reranking with bert.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021), 'Learning transferable visual models from natural language supervision', *Proceedings of Machine Learning Research* **139**, 8748–8763.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R. and Vasconcelos, N. (2010), A new approach to cross-modal multimedia retrieval, *in* 'Proceedings of the 18th ACM International Conference on Multimedia', ACM, pp. 251–260.

Robertson, S. and Zaragoza, H. (2009), 'The probabilistic relevance framework: Bm25 and beyond', *Foundations and Trends in Information Retrieval* **3**(4), 333–389.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. et al. (2022), 'Laion-5b: An open large-scale dataset for training next generation image-text models', *Advances in Neural Information Processing Systems* **35**, 25278–25294.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X. and Xu, C. (2022), Filip: Fine-grained interactive language-image pre-training, *in* 'International Conference on Learning Representations'.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C. et al. (2021), Florence: A new foundation model for computer vision, *in* 'arXiv preprint arXiv:2111.11432'.