

SemRank: Semantic Image Retrieval via Knowledge Graph Filtering and CLIP Embeddings

Sourav Pandey¹✉ and Ashish Singh Patel¹

National Institute of Technology Mizoram, Aizawl, Mizoram, 796012, India
{dt23cs011, ashish.cse}@nitmz.ac.in

Abstract. Image retrieval has become a critical component of modern information systems, driven by the growing need to search large-scale visual databases in domains such as e-commerce, healthcare, and digital media. Existing work image retrieval systems have shown the effectiveness in mapping text and visual into a shared embedding space, and allowing improved and fast image retrieval. However, these retrieval systems struggle with relational semantics, resulting in poor ranking performance and difficulty retrieving even the nearest-neighbour image that semantically matches the query. This work introduces SemRank, a hybrid retrieval framework that integrates Knowledge Graph-based semantic filtering with CLIP embeddings and FAISS similarity search to overcome these limitations. Moreover, a query subgraph is added to filter extracted entities and relationships from the query, eliminating semantically irrelevant candidates before neural ranking. CLIP embeddings then perform high dimensional semantic matching, accelerated through efficient FAISS indexing. Experiments are conducted on MS-COCO and Flickr30K, yielding results that show improvements over recent and baseline methods, achieving higher recall scores and reducing false positives, particularly for complex multi-object queries.

Keywords: Semantic search · Knowledge Graph · Query subgraph · Cross-model retrieval.

1 Introduction

Semantic image search seeks to retrieve relevant images by understanding the relationship between visual content and associated text. However, metadata in large datasets is often incomplete or noisy. Modelling image text co-occurrence helps refine and propagate semantic information, enabling better annotation of weakly labelled or unlabelled images and improving overall retrieval performance. The main challenge in semantic image retrieval involves two competing objectives: (1) achieving high computational efficiency to enable real-time retrieval on large-scale image databases (millions to billions of images), and (2) maintaining semantic accuracy such that retrieved images precisely match the user’s intent, including complex relationships and domain-specific contexts.

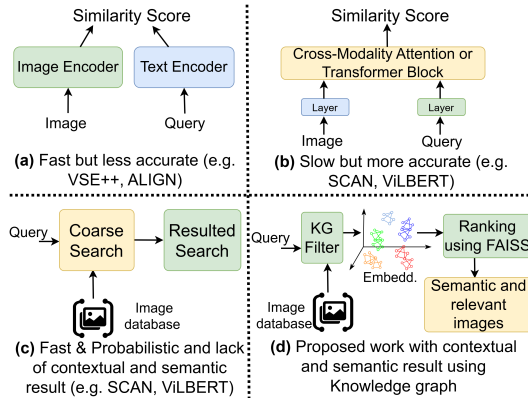


Fig. 1: Comparison of proposed work with the baseline and other related work

Traditional keyword-based and low-level visual feature methods struggle with complex semantic queries involving multiple entities and contextual relations [6, 8, 9]. Although deep learning and vision-language models improve cross-modal alignment through shared embeddings, enforcing explicit multi-entity constraints and relational consistency remains difficult in practice. Fig. 1 summarizes a comparative analysis of prior techniques and highlights these limitations. Prior methods exhibit complementary limitations such as two stage or probabilistic systems are efficient but cannot effectively enforce multi-entity relationships or fine-grained semantic constraints, which leads to high false-positive rates on complex queries [10, 21].

However, joint cross-modal encoders, such as cross-attention transformers, model detailed text-image interactions but cannot precompute similarities for large-scale approximate nearest-neighbour (ANN) indexing. Moreover, storing region-level features or intermediate activations for millions of images is prohibitively expensive in terms of memory and computation, rendering these approaches unsuitable for web scale retrieval.

To address these trade-offs, the proposed SemRank integrates Knowledge Graph (KG) filtered, CLIP embedding and FAISS (Facebook AI Similarity Search) [5] framework that unites symbolic reasoning, powerful vision language embeddings, and efficient vector search. The pipeline uses KG-based entity, and relation extraction to prune the candidate set to semantically coherent images, applies CLIP-derived embeddings for high-dimensional semantic matching, and leverages FAISS for fast ANN retrieval over precomputed vectors. By constraining the search space with explicit semantic filters, this approach preserves the performance advantages of transformer-based interaction. The main contributions of the paper are as follows:

1. Proposed semantic retrieval framework (SemRank) that integrates KG based symbolic reasoning with CLIP-based embeddings, enabling more accurate interpretation of natural language queries for image retrieval.

2. The query subgraph captures entities and relationships expressed in the query, improving graph consistency and relevance by filtering semantically irrelevant images and substantially reducing false positives before ranking.
3. To validate the proposed work ablation study conduct between in the Flickr30k and MS-COCO datasets, highlighting the impact of KG integration on semantic search and cross-modal retrieval performance.

The remainder of this article is organized as follows. Section 2 provides a brief review of related work. Section 3 introduces the proposed SemRank method and describes its architecture in detail. Section 4 outlines the experimental setup of the proposed approach. Section 5 presents the results and discussion, including comparisons with state-of-the-art retrieval methods, ablation studies, and visualization analyses. Finally, Section 6 concludes the article and highlights directions for future work.

2 Related Work

Despite rapid advances in vision language modelling, current image retrieval systems still face challenges in balancing semantic understanding with large-scale efficiency. Existing approaches can be broadly categorized into independent encoders, cross-modal attention models, and coarse-to-fine retrieval pipelines, each with distinct strengths and limitations.

Independent encoder methods represent the earliest scalable semantic retrieval systems. They employ separate image and text encoders to project both modalities into a shared embedding space, where relevance is computed using simple similarity measures. Representative models include VSE++ [6], ALIGN [21], CLIP dual-encoder, and Florence-2 embedding models, all trained using large-scale contrastive learning. While highly efficient, these approaches lack explicit cross-modal reasoning at inference time, making them ineffective for relational or compositional queries. Consequently, retrieval is often driven by superficial visual cues such as colour or dominant objects, and even strong models like CLIP achieve only moderate recall for complex multi-object queries without additional constraints.

Cross-modal attention and joint transformer architectures, including SCAN [10], ViLBERT [15], LXMERT [19], UNITER [3], OSCAR, and BLIP [11], enable fine-grained interaction between text tokens and visual regions, allowing reasoning over objects, spatial relationships, and complex semantics. However, these models cannot precompute similarity scores or efficiently leverage large-scale ANN indexing, resulting in prohibitively slow inference on million-scale image collections. Additionally, storing region-level features and intermediate activations incurs significant memory overhead [17].

Moreover, although more recent models such as BLIP-2 further improve vision language reasoning through integration of large language models, the proposed work uses CLIP-based dual encoders for large-scale retrieval efficiency.

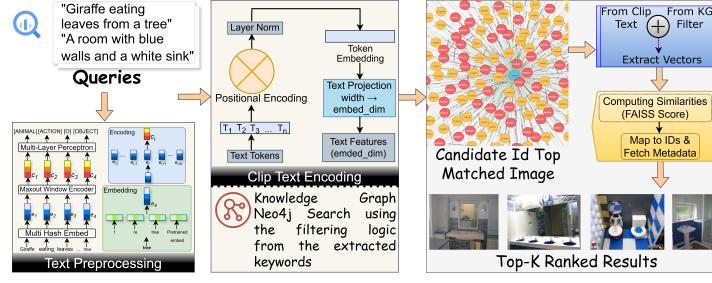


Fig. 2: The Proposed SemRank Architecture for the semantic search using the KG

Cross-attention architectures inherently prevent offline embedding precomputation required for efficient FAISS indexing, making them unsuitable for million-scale galleries. Therefore, CLIP enables fast similarity search through precomputed embeddings while still providing reasonably strong semantic alignment, offering a practical trade-off between relational reasoning and scalability.

Probabilistic retrieval models balance efficiency and accuracy through a coarse-to-fine strategy: fast initial filtering via keywords, visual features, or hashing, followed by expensive re-ranking using cross-attention on top candidates [8, 18]. Despite this, fundamental trade-offs persist, independent encoders scale well but lack relational semantics, while cross-modal attention models capture rich semantics but fail at scale, leaving early pipeline stages contextually weak. To address these limitations, the proposed KG-filtered CLIP-FAISS framework integrates symbolic semantic filtering with neural embeddings and ANN search, achieving both semantic precision and large-scale retrieval efficiency.

3 Methodology

This section presents SemRank, a framework for semantic search and retrieval of complex queries over large-scale image databases. The overall architecture is illustrated in Fig. 2, which depicts a complete framework comprising KG-based symbolic filtering, CLIP embeddings, and FAISS similarity search to retrieve the most semantically relevant images for a given natural query.

3.1 Text Preprocessing

Given an input string x (the query), using spaCy to extract the keywords and entities (K, E) through a sequence of functions. First tokenization splits x into a sequence of tokens $(t_1, \dots, t_n)$ where each t_i carries surface form w_i , lemma l_i , POS tag p_i and dependency label d_i . $Tok : X \rightarrow T^n, Tok(x) = (t_1, \dots, t_n)$. Statistical POS tagging and dependency parsing, then assign labels via conditional probabilities, producing a dependency tree over tokens as shown in Eq. 1.

$$p_i = \arg \max_{p \in \mathcal{P}} P(p \mid \text{context around } t_i), \quad (h_i, d_i) = \arg \max_d P(d \mid \text{context}). \quad (1)$$

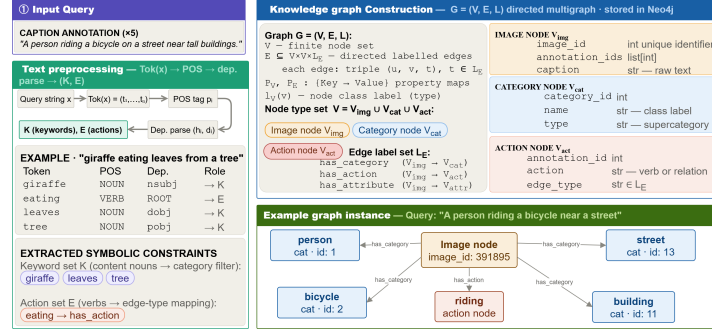


Fig. 3: Knowledge Graph construction framework in SemRank

Capturing content words like "giraffe", "eat", "leaf", "tree" from the query "giraffe eating leaves from a tree". These sets (K, E) serve as symbolic constraints for downstream Neo4j filtering and intent logging, bridging natural language to structured semantic search.

3.2 Clip Text Encoder

For semantic search, CLIP's text encoder converts the query into a numerical representation within a shared image text embedding space [22]. Given a tokenized query sentence with L tokens $(w_1, w_2, \dots, w_L)$ CLIP's text encoder maps it to a single embedding vector $e_i = E(w_i) \in \mathbb{R}^d$. Where $E : \mathcal{V} \rightarrow \mathbb{R}^d$ is the embedding matrix. Positional encoding p_i are added to form the input matrix $X = [e_1 + p_1, \dots, e_L + p_L]^T$. The sequence is processed by the transformer encoder, where each layer applies multi-head self-attention in Eq. 2.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where, Q , K and V are the query, key and value matrix thus followed by a feed-forward network and residual normalization, producing contextualized representations $H^{(N)} = [h_1, \dots, h_L]^T$. CLIP selects the final [EOS] representations as the raw sentence embedding $q_{raw} = h_{EOS}$ which is subsequently L2-normalized [17]. This normalized query vector serves as the semantic representation used during retrieval. For a database of candidate embeddings v_j that are also normalized, semantic similarity reduces to the inner product $s_j = q \cdot v_j = \cos(\theta_j)$ allowing CLIP [17] to rank items according to their cosine similarity with the query. Thus, the text encoder converts natural-language queries into numerical vectors that encode semantic content, enabling the retrieval of the most semantically similar images or items from the search space.

3.3 Knowledge Graph

The KG is constructed directly from the image dataset annotations. For each image in MS-COCO or Flickr30K [13, 16], a dedicated image node is created and

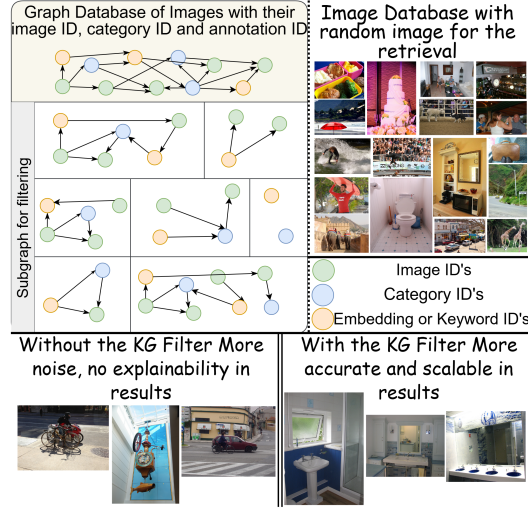


Fig. 4: Query subgraph representation within the KG for semantic image search

assigned a unique `image_id`. Category annotations linked to that image are extracted and instantiated as category nodes, each carrying a `category_id`, a class name (e.g., "person", "bicycle"), and a supercategory label. An action node is similarly created for every verb or relational phrase found in the associated captions, derived via the spaCy dependency. Directed edges of type `has_category` are drawn from each image node to its corresponding category nodes, and edges of type `has_action` connect the image node to its action nodes, forming a directed multigraph $G = (V, E, L)$. This construction is performed once as an offline preprocessing step and persisted in a Neo4j graph database, enabling fast symbolic querying at inference time. Fig. 3 illustrates the construction of the proposed KG.

The proposed architecture aims to address the gap in the recommendation of images from a large image dataset for fast and efficient retrieval using the graph database, while preserving graph structural consistency and relevance and improving robustness to noise and weak co-occurrence in the retrieval. For this problem, the proposed Algorithm 1 and the KG use a filtration process that improves the graph retrieval problem by introducing a query subgraph, which selects candidates on the basis of both the keywords and the embedded representation of the query. The KG is defined as a directed, multigraph $G = (V, E, L)$. Where, V is a finite set of nodes; $E \subseteq V \times V \times L_E$ is set of directed edges, each edge brings a triple (u, v, t) meaning an edge from node u to node v of type t , where t comes from a finite set of edge-labels L_E . Through the property mapping, each node is defined by a function $\mathcal{P}$ such that $\mathcal{P}_V; \mathcal{P}_E : \{(Key \rightarrow value) \text{ map}\}$ this assigns multiple edges of different types between the same pair of nodes, or even multiple edges of same type if allowed. And each node v has a label $l_V(v)$ identifying its class, and a property map $\mathcal{P}_V(v)$ storing attributes (e.g. for

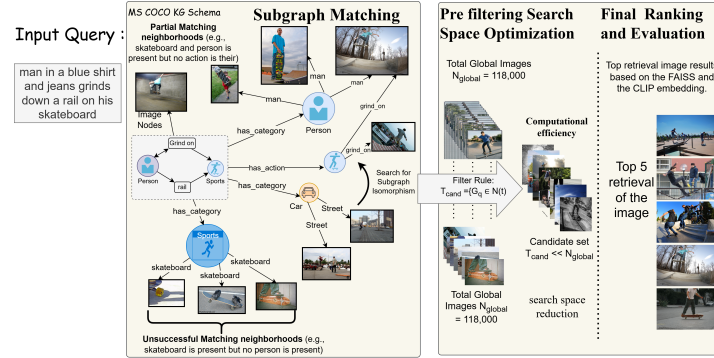


Fig. 5: Query Subgraph Workflow: illustrating the complete operation and foundational importance of the Query Subgraph (G_q) in proposed SemRank image retrieval framework

an image node: **image_id**, annotations and for a category node: **category_id**, type, ids). As categories and images scale, KGs face the following structural challenges:

1. **Density:** With many categories, the potential number of category-pairs is $\mathcal{O}(|V|^2)$. If many categories co-occur at least once, E may become very large, making storage, traversal, querying, and analysis expensive.
2. **Noise / weak co-occurrences:** Rare co-occurrences often don't reflect stable semantic relations, including them clutters the graph, lowers signal-to-noise ratio, and makes downstream tasks (e.g., clustering or recommendation) noisy or misleading.
3. **Structural Consistency:** Starting from a single graph node can retrieve semantically similar but structurally unrelated information, causing incoherent reasoning. Enforcing graph reachability constraints preserves structural consistency in semantic search results.
4. **Relevance:** Regardless of the retrieval method, result relevance is determined by a ranking algorithm that measures semantic similarity between the query and the content in the search graph, ensuring contextually meaningful and appropriate results.

When a user issues a query (for e.g., “find all images that contain a giraffe and a tree”), that query can be represented as a small graph call it the *query graph* (or query subgraph) $G_q = (V_q, E_q)$. Nodes in V_q correspond to entities that the query keywords match, and E_q matches the relationships that hold the same image and properties of the query. Thus G_q encodes a pattern (structure + constraints) that has to match against the full KG G . As shown in Fig. 5 complete workflow of the proposed subgraph querying reduces the complexity of the system by eliminating the irrelevant images from the database mathematically, subgraphs $g = (V', E')$ of G such that g matches the structure of the query graph G_q under the mapping

Algorithm 1 SemRank: Semantic Search with KG

Require: User query Q ; No. of retrieval results k ; KG $G = (V, E)$; Valid categories $\mathcal{C}_{valid} \subset V$; Pre-computed image embedd $\mathbf{E} \in \mathbb{R}^{N \times d}$; CLIP Model $\mathcal{M}$; Images $\mathbf{I}$.

Ensure: Ranked list of top- k images $\mathcal{R}$.

- 1: **Phase 1: Processing & Filtering**
- 2: $K_{raw} \leftarrow \text{NLPPProcess}(Q)$
- 3: $K_{fil} \leftarrow \{w \in K_{raw} \mid w \in \mathcal{C}_{valid}\}$
- 4: $\mathcal{I}_{cand} \leftarrow \emptyset$
- 5: **if** $K_{fil} \neq \emptyset$ **then** ▷ Filtering
- 6: $\mathcal{I}_{cand} \leftarrow \{i \in \mathbf{I} \mid \forall w \in K_{fil}, (i) \rightarrow (w) \in G\}$
- 7: **if** $\mathcal{I}_{cand} = \emptyset$ **then** ▷ Fallback
- 8: $\mathcal{I}_{cand} \leftarrow \{i \in \mathbf{I} \mid \exists w \in K_{fil}, (i) \rightarrow (w) \in G\}$
- 9: **end if**
- 10: **else**
- 11: $\mathcal{I}_{cand} \leftarrow \mathbf{I}(G)$
- 12: **end if**
- 13: **Phase 2: Semantic Similarity Ranking**
- 14: $\mathbf{v}_q \leftarrow \text{Encode}(Q, \mathcal{M})$ ▷ query embedd
- 15: $\mathbf{v}_q \leftarrow \frac{\mathbf{v}_q}{\|\mathbf{v}_q\|}$
- 16: $\mathcal{S} \leftarrow \emptyset$ ▷ score set
- 17: **for** each image $i \in \mathcal{I}_{cand}$ **do**
- 18: $\mathbf{v}_i \leftarrow \mathbf{E}[i]$
- 19: $\text{sim}(i, Q) \leftarrow \mathbf{v}_i \cdot \mathbf{v}_q$ ▷ Cosine similarity
- 20: $\mathcal{S} \leftarrow \mathcal{S} \cup \{(i, \text{sim}(i, Q))\}$
- 21: **end for**
- 22: $\mathcal{R} \leftarrow \text{SortDescend}(\mathcal{S})$
- 23: **return** $\mathcal{R}[1 : k]$

in Eq. 3:

$$f : V_q \rightarrow V' \subseteq V \quad (3)$$

Such that for every node $u \in V_q$, mapped node $f(u) \in V'$ has compatible properties (labels, constraints) as the specified in the query. And for every edge $(u, v) \in E_q$ there is a corresponding edge $(f(u), f(v)) \in E' \subseteq E$ in the data graph G , with matching relationship type and optionally matching edge properties. As in the Fig. 4 the given full KG with the nodes and the edges and the relevant query subgraph. For an e.g., “A room with blue walls and a white sink” the given full KG G the query engine will attempt to find subgraphs $g = (V', E') \subseteq G$ that are isomorphic to G_q under a mapping f . There is some image node $I \in V$ such that I (image) **CONTAINS** some nodes representing a “room” and these nodes have properties edge to a “Blue wall” attribute node. If all those relations hold in G , then I matches the query, and the subgraph containing those nodes and edges is returned as a result.

3.4 Semantic search

To provide accurate image recommendations for a query, a re-ranking stage is performed using the FAISS library [5]. These structured semantics are then mapped to a query subgraph, which is executed on the Neo4j KG to retrieve a candidate set of images whose metadata satisfies the extracted constraints and this graph. After filtering, embeddings capture nuanced semantic similarity beyond exact labels, enabling the retrieval of images that match context, style, composition, or other features implicitly encoded in embeddings. For instance, the semantic search for x query gives the ranked results, so mathematically in Eq. 4,

$$I = (K, E) = \{i \in V_{image} : \text{image } i \text{ have const. } \} \quad (4)$$

Here, V_{image} is the set of all image nodes in the KG and the query subgraph matches images whose connected category/attributes node match the constraints in (K, E) . As a result, $\mathcal{I} = \{i_1, i_2, \dots, i_N\}$ where N is less than the total number of images in the database and each image i_j matches the structured/filter constraints. And the set of query keywords be $K = \{k_1, k_2, \dots, k_m\}$ the framework first attempts to search keywords using algorithm 1. Finding the candidate set $\mathcal{C}_{candidate}$ is defined as the absolute intersection of the neighbour sets associated with the specific category nodes in Eq 5.

$$\mathcal{C}_{candidate} = \bigcap_{j=1}^m \{i \in \mathcal{V}_{image} \mid (i, v_{k_j}) \in \mathcal{E}\} \quad (5)$$

Therefore, when the user's query is overly restrictive or includes rare combinations of entities, the $\mathcal{C}_{candidate}$ intersection may yield an empty set ($\mathcal{C}_{strict} = \emptyset$). To ensure robustness and prevent framework failure, the architecture seamlessly falls back using the algorithm 1 to establish a set of fallback candidates $\mathcal{C}_{fallback}$ and is defined as Eq. 6:

$$\mathcal{C}_{fallback} = \bigcup_{j=1}^m \{i \in \mathcal{V}_{image} \mid (i, v_{k_j}) \in \mathcal{E}\} \quad (6)$$

Then, compute the query embedding $q = \text{CLIP_Text}(x) \in \mathbb{R}^d$. For the candidate images in $\mathcal{I}$ embeddings into a matrix in Eq. 7:

$$V_C = \begin{bmatrix} v_{i_1}^\top \\ v_{i_2}^\top \\ \vdots \\ v_{i_N}^\top \end{bmatrix} \in \mathbb{R}^{N \times d}; \|v_{i_j}\| = 1 \quad (7)$$

Compute similarity score vector in Eq. 8:

$$s = V_C q, s_j = v_{i_j} \cdot q = \cos(v_{i_j}, q) \quad (8)$$

Ranked list of image IDs with their similarity scores in Eq. 9:

$$\text{top_indices} = [(i_{j_1}, s_{j_1}), (i_{j_2}, s_{j_2}), \dots, (i_{j_5}, s_{j_5})] \quad (9)$$

4 Experiment

This section explains the experimental and dataset setup used for the construction of the proposed work.

4.1 Dataset

Flickr30K [16] and **MS-COCO** [13] are widely used for semantic search because they pair real images with natural, human written captions. Flickr30K contains about 31k everyday pictures with five short captions each. MS-COCO is much larger, with over 120k diverse and complex scenes, each captioned five times.

4.2 Experimental Setup

Evaluation Metrics To evaluate the performance of the proposed work, the *Recall@K* ($R@K = 1, 5$ and 10) is employed, which measures how well a semantic image retrieval system can find the correct image within its top- K search results [18].

Implementation Details The system interprets a user’s natural-language query using both KG entities and CLIP [17] embeddings. Neo4j serves as the symbolic filter, reducing the search space by selecting images connected through relevant objects, attributes, and relationships. CLIP *ViT-B/32* and *ViT-L/14* generate high-dimensional image-text embeddings, while FAISS (IndexFlatL2) enables fast and accurate vector similarity search [22]. All experiments were conducted on an NVIDIA RTX 4060 GPU (8 GB VRAM) with 48 GB RAM.

5 Result

This section presents detailed experimental results and an evaluation of the proposed SemRank semantic search.

5.1 Comparison with State-of-the-art Methods

Recent text-to-image retrieval methods have become popular, yet true semantic and contextual matching remains challenging especially at a large scale because many models still misinterpret visual features when the gallery grows. Early State-of-the-art (SOTA) models like SCAN [10] learns region word interactions and achieves strong Recall@K on COCO, but its similarity scoring is highly sensitive to partial matches, as the gallery size grows and overlapping regions introduce distractors and false positives due to the lack of explicit noise modelling. VSRN [12] improves semantics via region-level reasoning graphs, but relies heavily on object detector accuracy and clean region caption alignment, and

Table 1: Comparison of text to image retrieval performance on Flickr30K and MS-COCO datasets.

Methods	Flickr30K			MS-COCO		
	R@1	R@5	R@10	R@1	R@5	R@10
IMRAM [1]	16.6	41.8	54.1	39.6	69.1	79.8
VSRN [12]	25.5	47.6	58.6	40.5	70.6	81.1
SCAN [10]	35.5	65.0	75.2	38.6	69.3	80.4
SGR [4]	40.2	66.8	75.3	23.5	58.9	75.1
SAF [4]	49.7	73.6	78.0	57.8	86.4	91.9
NAAF [23]	52.5	79.3	86.2	58.5	86.0	91.3
GraDual [14]	60.4	86.7	92.0	65.3	91.9	96.4
CVSE [20]	61.9	87.1	92.0	59.9	89.4	95.2
GPO [2]	61.4	85.9	91.5	42.4	72.7	83.1
ESC [18]	59.1	83.8	89.1	64.8	90.7	96.0
MKG [7]	63.2	87.4	92.3	65.6	91.5	96.2
MKVSE [7]	63.2	87.4	92.3	66.4	92.1	96.6
Proposed Work (SemRank)	65.4	88.7	93.7	68.8	93.5	97.5

Table 2: Performance comparison of proposed work showing improvements from integrating the KG over the Baseline

Methods	Architecture	MS-COCO		
		R@1	R@5	R@10
(Without KG Filter)				
Baseline	Glove + ResNet18	25.0	30.1	42.2
Baseline	MiniLM + EfficientNet	37.8	53.5	62.3
(With KG Filter)				
Proposed work	CLIP ViT-B/32 + FAISS	63.7	87.0	90.1
Proposed work	CLIP ViT-L/14 + FAISS	68.8	93.5	97.5
Proposed work	CLIP ViT-B/32	35.0	70.0	74.4

with incomplete captions or beyond curated COCO settings, reasoning errors accumulate. SAF [4] mitigates weak alignments via vectorized similarity and filtration, yet remains a matching level approach and struggles with subtle noise or mismatched image caption pairs in large-scale retrieval.

Pooling-based methods like GPO [2] improve feature aggregation but do not resolve the deeper issue of noisy captions steering the embedding space incorrectly. More recent approaches, such as NAAF [23], model mismatched fragments as negative evidence, yet they mainly address local misalignment and cannot fully handle distribution-level noise. Moreover, most of these methods are evaluated only on small benchmarks (1K or 5K test splits), so their performance drops when moving to larger galleries where distractors increase and ranking heuristics overfit. Their pairwise matching and graph reasoning also scale poorly, forcing practical systems to rely on coarse pre-filtering. Fig. 6a shows retrieval examples from the proposed method, with green boxes indicating relevant images and red boxes marking irrelevant ones, while Table 1 reports SOTA comparisons on Flickr30k and MS-COCO [13].

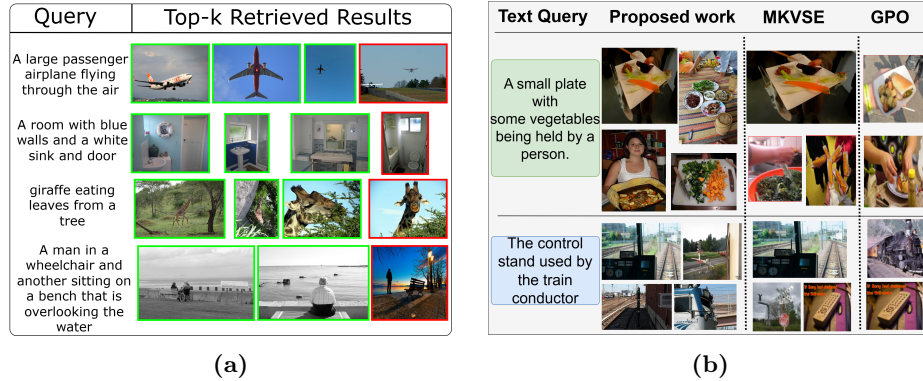


Fig. 6: (a) Semantic search results of the proposed method with the most relevant image and semantically/contextually rich images. (b) Qualitative comparison of the proposed work with MKVSE [7] and GPO [2] showing semantically and contextually relevant results on MS-COCO dataset.

5.2 Ablation Study

1. **Parameter Analysis and Impact of KG:** The baseline method, which combines GloVe word embeddings with ResNet18 visual features, achieves $R@1 = 25.0\%$, $R@5 = 30.1\%$, and $R@10 = 42.2\%$. In comparison, the MiniLM model with EfficientNet improves performance to $R@1 = 37.8\%$, $R@5 = 53.5\%$, and $R@10 = 62.3\%$. However, these scores remain modest, indicating that traditional cross-modal feature representations have limited semantic understanding and struggle to capture deeper image-text relationships. The full comparison is shown in Table 2. When KG filtering is combined with CLIP ViT-B/32, performance rises sharply to $R@1 = 63.7\%$, $R@5 = 87.0\%$, and $R@10 = 90.1\%$. The 38.7% improvement in $R@1$ represents a relative gain over the baseline, demonstrating the profound impact of KG-based filtering. This improvement comes from two effects: (1) the KG removes semantically irrelevant images, reducing false positives, and (2) the narrowed candidate set allows FAISS to focus on distinguishing only the most relevant matches. An ablation on CLIP [17] backbones shows that larger models give consistent gains. Comparing ViT-B/32 against ViT-L/14 reveals consistent improvements. ViT-B/32 reaches $R@1 = 63.7\%$, $R@5 = 87.0\%$, $R@10 = 90.1\%$, while ViT-L/14 improves to $R@1 = 68.8\%$, $R@5 = 93.5\%$, $R@10 = 97.5\%$, with increases of 5.1, 5.3, and 7.1 percentage points, respectively. The higher patch resolution of ViT-L/14 helps capture more detailed visual cues, resulting in more discriminative embeddings. These monotonic gains indicate that larger Vision Transformers better capture semantic nuance, especially when paired with KG-based candidate filtering. Notably, the larger boost in $R@10$ suggests that increased model capacity is particularly useful when multiple correct candidates exist, as richer embeddings help rank them more effectively.

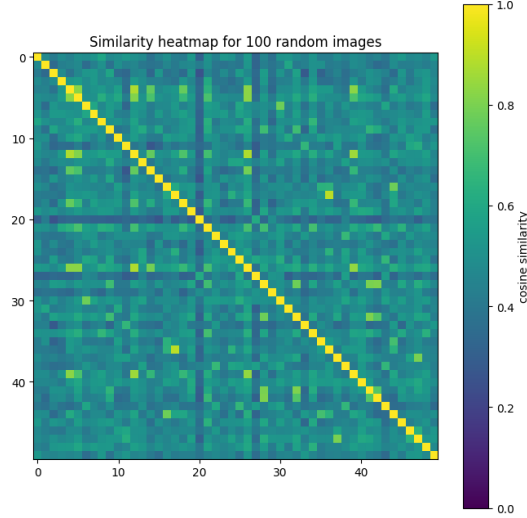


Fig. 7: Results of the proposed method with the most relevant image with semantic and contextually rich images

2. **Qualitative Results:** The qualitative analysis (Fig. 6b) shows the proposed KG framework yields more semantically accurate retrievals than MKVSE [7] and GPO [2] on two test queries. For e.g., *A small plate with some vegetables being held by a person*, the KG correctly enforces object and relation constraints (plate, vegetables, person, containing/holding). MKVSE returns plates and vegetables but often omits the person holding relation, GPO returns visually similar but semantically wrong results (e.g., hands with paint/dye mistaken for vegetables). For *The control stand used by the train conductor*, the KG correctly grounds the specialized, domain-specific object in the train context. MKVSE [7] captures general train scenes but not the specific control stand, GPO exhibits semantic drift, returning unrelated items like stop signs.
3. **Pairwise Cosine Similarity Analysis of Candidate Images:** The pairwise cosine similarity heatmap helps in explaining the logical retrieval and semantic evaluation. While KG ensures that all candidates in the matrix strictly satisfy the query constraints ($G_q \subseteq \text{Neighbourhood}(i)$), the heatmap (shown in Fig. 7) reveals the topological distribution of these candidates in the CLIP [17] latent space. If a system returns 10 images that are nearly identical (e.g., consecutive frames from the same video or duplicated data), $R@k$ might be artificially high, but user utility is low. The heatmap visualizes a matrix heavily dominated by scores near 1.0, indicating high redundancy. Conversely, a matrix with varied scores indicates a healthy diversity of visual representations for the same query. Moreover, within the top- k results, images often group into distinct "sub-concepts". For example, if the query is

"person on a bicycle", the heatmap might reveal two distinct high-similarity clusters (blocks of warm colours): one representing mountain biking and another representing city commuting. Therefore, the heatmap shows that the CLIP model successfully captures fine-grained contextual nuances.

6 Conclusion

This work presents a unified framework (SemRank) for retrieving the semantically rich images for a large dataset, with the most appropriate image from complex queries, and improves semantic understanding and large-scale efficiency. By integrating KG based symbolic filtering with CLIP embeddings and FAISS vector search, the approach enforces multi-entity relational constraints while preserving the efficiency of dual-encoder architectures. Experimental results demonstrate substantial improvements over traditional baselines, highlighting the importance of combining structured reasoning with modern vision language representations. The framework not only reduces false positives but also scales efficiently to large databases, making it a practical and robust solution for real-world semantic search applications where semantic and context relevant text to image is required. Future work will explore dynamic KG construction, multimodal grounding, and adaptive retrieval mechanisms to further enhance contextual understanding for more complex queries.

Supplemental Material Statement:

The source code for the architecture is available on the repository at the following link: <https://github.com/Sourav544/SemRank-Semantic-search.git>

References

1. Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12655–12663 (2020)
2. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15789–15798 (2021)
3. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
4. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1218–1226 (2021)
5. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. *IEEE Trans. Big Data* (2025)
6. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017)

7. Feng, D., He, X., Peng, Y.: Mkvse: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* **19**(5), 1–21 (2023)
8. Han, J., Ngan, K.N., Li, M., Zhang, H.: Learning semantic concepts from user feedback log for image retrieval. In: 2004 IEEE International Conference on Multimedia and Expo (ICME). vol. 2, pp. 995–998. IEEE (2004)
9. Huang, H., Nie, Z., Wang, Z., Shang, Z.: Cross-modal and uni-modal soft-label alignment for image-text retrieval. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 18298–18306 (2024)
10. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 201–216 (2018)
11. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International conference on machine learning*. pp. 12888–12900. PMLR (2022)
12. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4654–4662 (2019)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *European conference on computer vision*. pp. 740–755. Springer (2014)
14. Long, S., Han, S.C., Wan, X., Poon, J.: Gradual: Graph-based dual-modal representation for image-text matching. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 3459–3468 (2022)
15. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **32** (2019)
16. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2641–2649 (2015)
17. Schall, K., Barthel, K.U., Hezel, N., Jung, K.: Optimizing clip models for image retrieval with maintained joint-embedding alignment. In: *International Conference on Similarity Search and Applications*. pp. 97–110. Springer (2024)
18. Shi, H., Liu, M., Mu, X., Song, X., Hu, Y., Nie, L.: Breaking through the noisy correspondence: A robust model for image-text matching. *ACM Transactions on Information Systems* **42**(6), 1–26 (2024)
19. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019)
20. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: *European conference on computer vision*. pp. 18–34. Springer (2020)
21. Wehrmann, J., Kolling, C., Barros, R.C.: Adaptive cross-modal embeddings for image-text alignment. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 34, pp. 12313–12320 (2020)
22. Yuan, X., Li, J., Xia, W., Gan, W., Ying, M., Hua, L.: Adapting cross-modal semantic discrepancy in text-based person search. 2025 IEEE International Conference on Multimedia and Expo (ICME) pp. 1–6 (2025)
23. Zhang, K., Mao, Z., Wang, Q., Zhang, Y.: Negative-aware attention framework for image-text matching. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 15661–15670 (2022)